DEEP LEARNING BASED FACE IMAGE SYNTHESIS

by

Xing Di

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

August, 2021

Abstract

Face image synthesis is an important problem in the biometrics and computer vision communities due to its applications in law enforcement and entertainment. In this thesis, we develop novel deep neural network models and associated loss functions for two face image synthesis problems, namely thermal to visible face synthesis and visual attribute to face synthesis.

In particular, for thermal to visible face synthesis, we propose a model which makes use of facial attributes to obtain better synthesis. We use attributes extracted from visible images to synthesize attribute-preserved visible images from thermal imagery. A pre-trained attribute predictor network is used to extract attributes from the visible image. Then, a novel multi-scale generator is proposed to synthesize the visible image from the thermal image guided by the extracted attributes. Finally, a pre-trained VGG-Face network is leveraged to extract features from the synthesized image and the input visible image for verification.

In addition, we propose another thermal to visible face synthesis method based on self-attention generative adversarial network (SAGAN) which allows efficient attention-guided image synthesis. Rather than focusing only on synthesizing visible faces from thermal

ABSTRACT

faces, we also propose to synthesize thermal faces from visible faces. Our intuition is based on the fact that thermal images also contain some discriminative information about the person for verification. Deep features from a pre-trained Convolutional Neural Network (CNN) are extracted from the original as well as the synthesized images. These features are then fused to generate a template which is then used for cross-modal face verification.

Regarding attribute to face image synthesis, we propose the Att2SK2Face model for face image synthesis from visual attributes via sketch. In this approach, we first synthesize facial sketch corresponding to the visual attributes and then generate the face image based on the synthesized sketch. The proposed framework is based on a combination of two different Generative Adversarial Networks (GANs) (1) a sketch generator network which synthesizes realistic sketch from the input attributes, and (2) a face generator network which synthesizes facial images from the synthesized sketch images with the help of facial attributes.

Finally, we propose another synthesis model, called Att2MFace, which can simultaneously synthesize multimodal faces from visual attributes without requiring paired data in different domains for training the network. We introduce a novel generator with multimodal stretch-out modules to simultaneously synthesize multimodal face images. Additionally, multimodal stretch-in modules are introduced in the discriminator which discriminate between real and fake images.

Primary Reader and Advisor: Prof. Vishal M. Patel

Secondary Reader: Prof. Rama Chellappa and Prof. Carlos Castillo

Acknowledgments

First of all, I would like to express the gratitude to my advisor Prof. Vishal Patel from the bottom of my heart for his support, encouragement and supervision during my Ph.D. program. Prof. Patel lead me to the field of facial recognition, image synthesis and generative model. He also provided me with excellent research environment where lots of interesting ideas come up during our discussions. Besides, I benefited significantly from his demanding research altitude, precise time management and insightful thoughts. Additionally, I got lots of well-trained skills like research, writing, communication, presentation under his guidance. None of these achievements during my Ph.D. program would have been possible without him.

Besides, I would like to thank other committee members of my dissertation: Prof. Rama Chellappa and Prof. Carlos Castillo for their valuable suggestions to my dissertation. Additionally, I would greatly thank Prof. Wei Shen, Prof. Najim Dehak, Prof. Carey Priebe and Prof. Alan Yuille for serving as my GBO exam members. It is my fortune to have each of them serving as my committee.

I would also thank my internship mentors Prof. Bo Wang (University of Toronto) and

ACKNOWLEDGMENTS

Dr. Dazhou Guo, Dr. Dakai Jin and Dr. Le Lu at PAII. Inc. They all spent lots of effort on helping me solving various problems and difficulties. In addition, I would thank the other co-authors, He Zhang, Vishwanath Sindagi, Shuowen Hu (U.S. Army Research Laboratory), Nathaniel Short (Booz Allen Hamilton), Benjamin S Riggan (University of Nebraska) for their kind help and valuable discussion during my research.

Many thanks to my colleagues in the Electrical Engineering Department in Johns Hopkins University and Rutgers University. The people there made my Ph.D. time colorful and memorable.

To the end, special thank to my wife Jingwan Fu and my family. I cannot image how to survive through my Ph.D. life without their encouragement and support.

Dedication

To my son Louis, who messed up my life in the COVID-19 pandemic.

Contents

List of Tables

LIST OF TABLES

List of Figures

Chapter 1

Introduction

1.1 Motivation

Face image synthesis is a widely studied problem in the computer vision and biometric research communities due to its applications in forensic, entertainment, and security. Face image synthesis studies the problem of how to train a generative model to generate facial samples. In particular, given a large collection of facial image data, a generative model is trained to generate diverse photo-realistic samples that match the given observed data distribution. Extracting meaningful representation of the data is significantly important to synthesize the data distribution. Recently, deep learning-based generative models have shown remarkable progress in data synthesis. Variational Autoencoder (VAE) [5, 6] is designed to solve this problem by maximizing the lower bound of data likelihood. Autoregressive models [7, 8] utilize the deep neural network to model the pixel space for

generating high-quality images.

Generative adversarial networks (GANs) [9] are another class of deep generative models that are utilized to synthesize realistic images by effectively learning the observed image distribution. A GAN consists of two parts – a generator and a discriminator that are trained iteratively. The discriminator is trained to distinguish between real samples from the true data distribution and the fake samples produced from the generator. The advances of GANs are explored in various applications like attribute manipulation [10–12], face frontalization [13–15], and super-resolution [16, 17].

In this thesis, the following three face image synthesis problems are studied. Generating high-resolution facial images from visual attributes is an extremely difficult problem because the model is required to learn the mapping from a very semantically abstract space to a diverse and complex RGB image space. This task requires the generated images to be not only realistic but also semantically meaningful. Existing generative models [18–20] either fail in generating photo-realistic images or produce images which don't match the attributes that are used to synthesize images. In addition, generating high-resolution facial images is difficult. Simply adding more upsampling layers usually leads to unstable training and meaningless outputs [21, 22].

Generating multimodal facial images from visual attributes is another challenging problem which we address in this thesis. Collecting paired multimodal face images is very expensive and difficult [23, 24]. A model that can simultaneously generate paired multimodal face images based on the visual attributes is significant in many practical scenarios like

forensics, entertainment and law assistant. A naive solution to this problem is to simply use an attribute-to-face synthesis model and then use another multimodal image-to-image translation model. However, these two kinds of models are extremely difficult problems themselves and the combination will not be an effective approach [25].

In addition, GANs are also employed for heterogeneous face recognition. To verify face images in two different modalities (i.e. thermal and visible), image-to-image translation models are utilized to bridge the domain gap. Although existing approaches [26–28] are capable to generate plausible facial images, the synthesized results are still far from optimal. One common issue is the loss of semantic attribute information such as expression, facial hair, gender, etc during synthesis. Such reconstruction degrades the performance of thermal-to-visible heterogeneous face verification.

1.2 Contributions

To address these challenges and issues, we conduct extensive research on the design of novel neural network architectures, modifying training objective functions, adding training regularization and strategies. The related publications are as follows:

- ”Facial Synthesis From Visual Attributes via Sketch Using Multiscale Generators,” in IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 2, no. 1, pp. 55-67, Jan. 2020.

- ”Multi-Scale Thermal to Visible Face Verification via Attribute Guided Synthesis,”

in IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3, no. 2, pp. 266-280, April 2021.

- "Multimodal Face Synthesis from Visual Attributes." in IEEE Transactions on Biometrics, Behavior, and Identity Science, doi: 10.1109/TBIOM.2021.3082038.

The main contributions for each work can be summarized as follows:

(1) New multi-scale generators are proposed for facial image synthesis from visual attributes via sketches. First, we formulate the attribute-to-face generation problem as a stage-wise learning problem, i.e., attribute-to-sketch, and sketch-to-face. Then, we propose a novel visual attribute conditioned sketch-to-face synthesis network for the face generator. The network is composed of an attribute augmentation module and a UNet [29] shape translation network. With the help of the attribute-augmentation module, the training stability is improved and the generators are able to synthesize diverse set of realistic face/sketch images. Finally, extensive experiments are conducted on the CelebA dataset [30] and the LFWA dataset [31] to demonstrate the effectiveness of the proposed image synthesis method. Furthermore, an ablation study is conducted to demonstrate the improvements obtained by different stages of our framework.

(2) A novel end-to-end multi-scale GAN is designed for thermal to visible face verification via attribute guided synthesis. The Multi-AP-GAN is developed for synthesizing high-quality visible faces from thermal images guided by facial attributes. A single generator with multi-scale output architecture and a Multimodal Compact Bilinear (MCB) pooling module [32, 33] are used to generate high-quality visible images. A novel triplet-

pair discriminator is proposed, where the discriminator [19] not only learns to discriminate between real/fake images but also discriminate between images/visual-attributes. Extensive experiments are conducted on three different volumes of the ARL Multimodal Facial Database [1, 34, 35] as well as the Thermal and Visible Paired Face Database [36] and the TUFTS face dataset [37]. Comparisons are performed against several recent state-of-the-art approaches. Furthermore, an ablation study is conducted to demonstrate the improvements obtained by including semantic attribute information for synthesis.

(3) A novel GAN that can simultaneously synthesize multimodal face images from visual attributes is designed. We introduce multimodal stretch-out and stretch-in modules in the generator and discriminator networks, respectively. In addition, a progressive training strategy is employed to generate multimodal photo-realistic high resolution images with consistant identity. Extensive experiments are conducted on three datasets: CelebA-HQ [30, 38] dataset, ARL Multimodal Face dataset [1, 34, 35] and CASIA-NIR-VIS 2.0 [39] to demonstrate the effectiveness of the proposed multimodal image synthesis method. To the best of our acknowledge, this is the first approach for synthesizing high-quality multimodal face images from visual attributes.

1.3 Dissertation Outline

The remaining thesis is organized as follows. Some core concepts of Generative Adversarial Networks (GANs) and variations of GANs are introduced in Chapter 2. Then,

Facial Synthesis From Visual Attributes via Sketch Using Multiscale Generator network is proposed in Chapter 3. The Multi-Scale Thermal to Visible Face Verification via Attribute Guided Synthesis work is introduced in Chapter 5. Then, Multimodal Face Synthesis from Visual Attributes work is introduced in Chapter 4. Some other related published works are introduced in Chapter 6, Chapter 8 and Chapter 7. Finally, conclusion and discussion about the future work are presented in Chapter 9.

Chapter 2

Background

In this chapter, we give a review on the literature about the deep generative networks and deep face image synthesis. In particular, we give a brief background of generative models, visual description-based synthesis and thermal-to-visible face synthesis problems. In addition, some recent face frontalization works are also reviewed.

2.1 Deep Generative Networks

Recent advances in deep learning have led to the development of various deep generative models for the problem of text-image synthesis and image-to-image translation [5–7, 9, 40–50]. Among them, variational autoencoder (VAE) [5, 6, 42], generative adversarial network (GAN) [9, 40, 45–48, 50], and Autoregression [7] are the most widely used approaches.

VAEs [5,6] are powerful generative models that use deep networks to describe distribution of observed and latent variables. A VAE model consists of two parts, with one network encoding a data sample to a latent representation and the other network decoding latent representation back to data space. VAE regularizes the encoder by imposing a prior over the latent distribution. Conditional VAE (CVAE) [18, 41] is an extension of VAE that models latent variables and data, both conditioned on a side information such as a part or label of the image. However, due to the imperfect element-wise square error measurements, the VAE model usually generates blurry images [51].

GANs [9] are another class of generative models that are used to synthesize realistic images by effectively learning the distribution of training images [15, 52, 53]. The goal of GAN is to train a generator G, to produce samples from training distribution such that the synthesized samples are indistinguishable from actual distribution by the discriminator, D. Conditional GAN is another variant where the generator is conditioned on additional variables such as discrete labels, text or images. The objective function of a conditional GAN is defined as follows

$$\begin{aligned} L_{cGAN}(G, D) = E_{\mathbf{x},\mathbf{y}\sim P_{data}(\mathbf{x},\mathbf{y})}[\log D(\mathbf{x}, \mathbf{y})]+ \\ E_{\mathbf{x}\sim P_{data}(\mathbf{x}),\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(\mathbf{x}, G(\mathbf{x}, \mathbf{z})))], \end{aligned} \tag{2.1}$$

where $\mathbf{z}$ the input noise, $\mathbf{y}$ the output image, and $\mathbf{x}$ the observed image, are sampled from distribution $P_{data}(\mathbf{x}, \mathbf{y})$ and they are distinguished by the discriminator, D. While for the generated fake $G(\mathbf{x}, \mathbf{z})$ sampled from distributions $\mathbf{x} \sim P_{data}(\mathbf{x}), \mathbf{z} \sim p_z(\mathbf{z})$ would like to

fool D.

2.2 Attribute-based Image Synthesis

Various models have been proposed in the literature for visual attribute/text description-based synthesis. For example, Yan *et al.* [18] proposed a discomposing conditional VAE (disCVAE) model to synthesize facial images from visual attributes. They took the assumption that a face image could be decomposed into two parts: foreground and background. By taking this assumption, the disCVAE model is able to generate plausible face images with corresponding attributes. Isola *et al.* [54] proposed Conditional GANs [55] for several tasks such as labels to street scenes, labels to facades, image colorization, etc. Built on this, Reed *et al.* [19] proposed a conditional GAN network to generate images conditioned on the text description. Several text-to-image synthesis works have been proposed in the literature that make use of the multi-scale information [20–22, 43, 56–58]. Zhang *et al.* [20] proposed a two-stage stacked GAN (StackGAN) method which achieves the state-of-the-art image synthesis results. More recently this work was extended in [21] by using additional losses and better fine-tuning procedures. Xu *et al.* [59] proposed an attention-driven method to improve the synthesis results. Zhang *et al.* [22] (HDGAN) adopted a multi-adversarial loss to improve the synthesis by leveraging more effective image and text information at multi-scale layers.

2.3 Heterogeneous Face Synthesis

Synthesis-based thermal-to-visible face verification algorithms leverage the synthesized visible faces for verification. Due to the success of CNNs and recently introduced GANs in synthesizing realistic images, various deep learning-based approaches have been proposed in the literature for thermal-to-visible face synthesis [26–28, 60–62]. For instance, Riggan *et al.* [27] proposed a two-step procedure (visible feature estimation and visible image reconstruction) to solve the cross-modal verification problem. Zhang *et al.* [26] proposed an end-to-end GAN-based approach for synthesizing photo-realistic visible face images from the corresponding polarimetric thermal images. Recently, Riggan *et al.* [28] proposed a new synthesis method to enhance the discriminative quality of generated visible face images by leveraging both global and local facial regions. Zhang *et al.* [34] introduced a multi-stream fusion-based generative model for cross-modal face verification. Di *et al.* [63] proposed a GAN-based network called AP-GAN to improve the synthesized visible image by utilizing visual attributes. Di *et al.* [64] proposed another unsupervised generative model which combines features from both thermal-to-visible and visible-to-thermal synthesized images for heterogeneous face verification. Recently Pereira *et al.* [65] proposed a generic adaptation-based network for heterogeneous face recognition. He *et al.* [61] proposed a generative model for thermal-to-visible face synthesis by utilizing texture inpainting and pose correction. Another improved FusionNet was proposed in [66], which increases robustness against overfitting using dropout for a thermal-to-visible generation. This method was evaluated on the RGB-D-T dataset [67]. Recently, Mallat *et al.* [68, 69] proposed a

cascaded model which is optimized by the contextual loss [70] for cross-spectrum synthesis. An attribute-guided visible face synthesis method using a conditional CycleGAN framework was proposed in [71].

2.4 Face Frontalization

Recent face frontalization methods utilize either 2D/3D warping [72,73,73–75], stochastic modeling [76–78] or the generative models [13–15, 79–85]. For instance, Hassner *et al.* [72] proposed a single unmodified 3D surface model for frontalization. Sagonas *et al.* [76] proposed a model that jointly performs frontalization and landmark localization by solving a low-rank optimization. Rui *et al.* [15] developed TP-GAN with a two-path architecture for capturing both the global and local contextual information. Tran *et al.* [80] introduced the DR-GAN model which frontalizes face images via disentangled representations. Hu *et al.* [14] introduced CAPG-GAN to synthesize frontal face images guided by the target landmarks. Zhao *et al.* [82] proposed the PIM model based on the domain adaptation strategy for pose invariant face recognition. Cao *et al.* [79] proposed the HF-PIM model which includes dense correspondence field estimation and facial texture map recovery. Similarly, Zhang *et al.* [84] developed the A3F-CNN model with the appearance flow constrain. Li *et al.* [86] proposed a frontalization model using a series of discriminators optimized by segmented face images. Yin *et al.* [13] proposed a self-attention-based generator is to integrate local features with their long-range dependencies for obtaining better

frontalized faces. Wei *et al.* [83] proposed the FFWM model to overcome the illumination issue via flow-based feature warping.

Chapter 3

Facial Synthesis From Visual Attributes via Sketch Using Multiscale Generators

Facial attributes are descriptions or labels that can be given to a face by describing its appearance [87]. In the biometrics community, attributes are also referred to as soft-biometrics [88]. Various methods have been developed in the literature for predicting facial attributes from images [89–95]. In this work, we aim to tackle the inverse problem of synthesizing faces from their corresponding attributes (see Fig. 3.1). Visual description-based facial synthesis has many applications in law enforcement and entertainment. For example, visual attributes are commonly used in law enforcement to assist in identifying suspects involved in a crime when no facial image of the suspect is available at the crime scene. This is commonly done by constructing a composite or forensic sketch of the person based on the visual attributes.

(a) (b)

Figure 3.1: Attribute prediction vs. face synthesis from attributes. (a) Attribute prediction: given a face image, the goal is to predict the corresponding attributes. (b) Face synthesis from attributes: given a list of facial attributes, the goal is to generate a face image that satisfies these attributes.

Reconstructing an image from attributes or text descriptions is an extremely challenging problem because the model is required to learn the mapping from a semantic abstract space to a complex RGB image space. This task requires the generated images to be not only realistic but also semantically consistent, i.e., the generated face images should preserve the facial structure as well as the content described in attributes. Several recent works have attempted to solve this problem by using recently introduced CNN-based generative models such as conditional variational auto-encoder (CVAE) [5, 18, 41] and generative adversarial network (GAN) [9,19–22,59]. For instance, Yan *et al.* [18] proposed a disentangled CVAE-based method for attribute-conditioned image generation. In a different approach, Reed *et al.* [19] introduced a GAN-based method for synthesizing images from detailed text descriptions. Similarly, Zhang *et al.* [20] proposed the StackGAN method for synthesizing photo-realistic images from text.

It is well-known that CVAE-based methods often generate blurry images due to the injected noise and imperfect element-wise squared error measure used in training [51]. In

contrast, GAN-based methods have shown to generate high-quality images [9]. In order to synthesize photo-realistic facial images, rather than directly generating an image from attributes, we first synthesize a sketch image corresponding to the attributes and then generate the facial image from the synthesized sketch. Our approach is motivated by the way forensic sketch artists render the composite sketches of an unknown subject using a number of individually described parts and attributes. Our approach is also inspired by the recent works [41, 96–99] that have shown the effectiveness of stage-wise training.

3.1 Proposed Method

In this chapter, we provide details of the proposed GAN-based attribute to face synthesis method, which consists of two components: sketch generator and face generator. Note that the training phase of our method requires ground truth attributes and the corresponding sketch and face images. Furthermore, the attributes are divided into two separate groups - one corresponding to texture and the other corresponding to color. Since sketch contains no color information, we use only the texture attributes in the first component (i.e. sketch generator) as indicated in Fig. 5.1.

In order to explore the multi-scale information during training, inspired by the previous works [20–22, 100], we adopt the idea of hierarchically-integrated multiple discriminators at different layers in our generators. The sketch/face generator network learns the training data distribution from low-resolution to high-resolution. This also helps in improving the

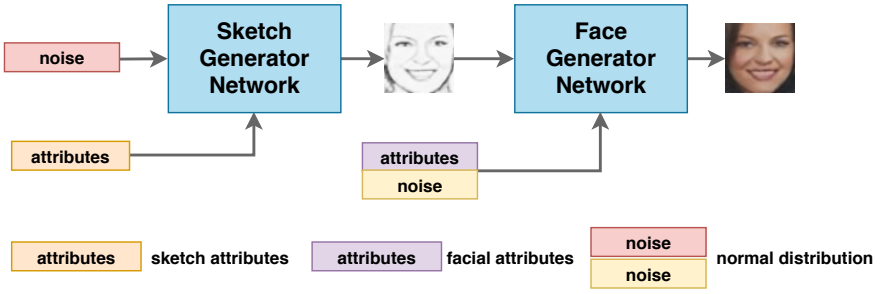


Figure 3.2: An overview of the proposed synthesis method. Given a noise vector sampled from the normal distribution, the Sketch Generator Network synthesizes sketch image conditioned on the sketch attributes. The synthesized sketch is then given as an input to the Face Generator Network, which outputs the high-quality face images conditioned on the facial attributes.

training stability of the overall network [101].

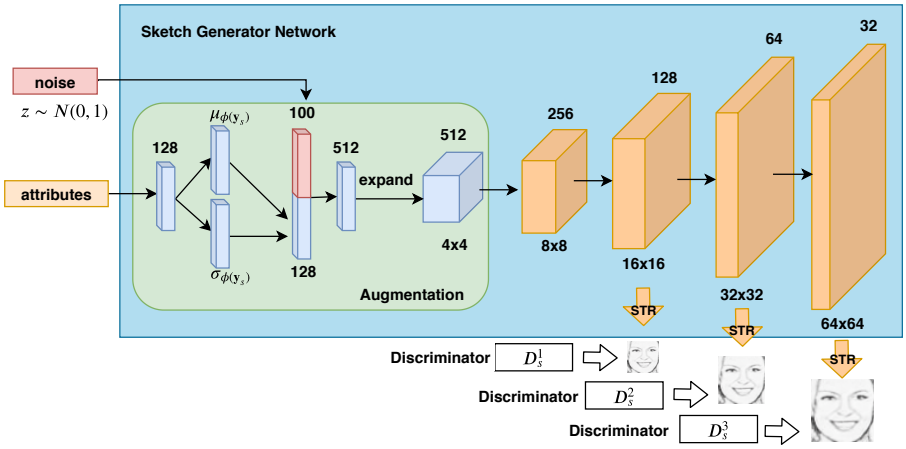


Figure 3.3: The sketch generator network architecture. The sketch attributes are first augmented by the attribute augmentation module, in which a new latent attribute variable is re-sampled from the estimated latent distribution ($\mu_{\phi(\mathbf{y}_s)}$ and $\sigma_{\phi(\mathbf{y}_s)}$) and concatenated with a noise vector. Then, the remaining up-sample modules (orange) aim to generate a series of multi-scale sketches with the augmented sketch attributes.

Stage 1: Attribute-to-Sketch

An overview of the sketch generator network architecture is shown in Fig. 3.3. Given the sketch attribute vector $\mathbf{y}_s$, the goal of the sketch generator network G_s is to produce multi-scale sketch outputs as follows

$$G_s(\mathbf{z}_s, \mathbf{y}_s) = \{\hat{\mathbf{x}}_s^1, \hat{\mathbf{x}}_s^2, \cdots, \hat{\mathbf{x}}_s^m\} \triangleq \hat{\mathbf{X}}_s, \tag{3.1}$$

where $\mathbf{z}_s$ is the noise vector sampled from a normal Gaussian distribution, $\{\hat{\mathbf{x}}_s^1, \hat{\mathbf{x}}_s^2, \cdots, \hat{\mathbf{x}}_s^m\}$ are the synthesized sketch images with gradually growing resolutions, and $\hat{\mathbf{x}}_s^m$ is the final output with the highest resolution. In order to explore the multi-scale information at different image resolutions, a set of distinct discriminators $D_s = \{D_s^1, ..., D_s^m\}$ are implemented for each $\hat{\mathbf{x}}_s^i, i = 1, 2, 3, \cdots, m$. An example of 3-scale generator architecture is shown in Fig. 3.3. It can be observed that the output sketch images are generated from the feature maps with certain resolutions (width $\times$ height) from different layers of the network.

The generator network consists of three modules: the attribute augmentation module (AA), the up-sample module (UP), and the stretching module (STR). The STR module consists of two 1×1 convolution layers followed by a Tanh layer, which aims to convert the feature map into a 3-channel output image. The UP module consists of an up-sampling layer followed by convolutional, batch normalization, and ReLU layers. Between each UP module, there is an additional residual block (Res) module [102, 103].

The AA module consists of a series of fully-connected neural networks which aim to

learn a latent representation of the given visual attribute vector $\mathbf{y}$. During training, we randomly sample a latent variable $\hat{\mathbf{y}}$ from an independent Gaussian distribution $\mathcal{N}(\mu_{\phi(\mathbf{y})}, \sigma_{\phi(\mathbf{y})})$, where the mean $\mu_{\phi(\mathbf{y})}$ and the diagonal covariance matrix $\sigma_{\phi(\mathbf{y})}$ are learned as the functions of visual attributes $\mathbf{y}$. In order to avoid over-fitting, the following KL-divergence regularization term is added during training between the augmented visual attribute distribution and the standard Gaussian distribution

$$\mathcal{L}_{aug} = \mathcal{D}_{KL}(\mathcal{N}(\mu_{\phi(\mathbf{y})}, \sigma_{\phi(\mathbf{y})}) \| \mathcal{N}(0, \mathcal{I})), \tag{3.2}$$

where $\mathcal{N}(0, \mathcal{I})$ is the normal Gaussian distribution [5, 20, 42]. Different from previous works [20, 21], we replace the traditional batch normalization layers with the conditional batch normalization [104] in order to overcome the attribute vanishing problem.

As shown in Fig. 3.3, the overall sketch generator network architecture is as follows:
AA(512)-UP(256)-Res(256)-UP(128)Res(128)-UP(64)-Res(64)-UP(32),
where the number in round bracket indicates the output channel of feature maps. As shown in Fig. 3.3, the three stretching (STR) modules convert the feature maps into 3-channel output sketch images at different resolutions.

Discriminator and Training Loss: The proposed sketch generator produces multi-scale resolution synthesized sketch images. In order to leverage the hierarchical property of the network, a set of discriminators $D_s = \{D_s^1, ..., D_s^m\}$ with similar architectures are designed for each scale. For a particular scale, the sketch discriminator is developed as

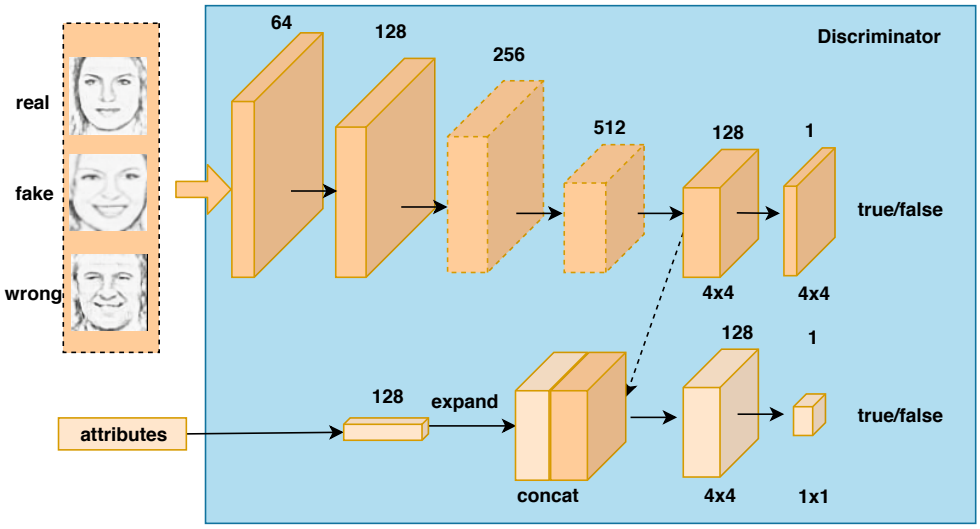


Figure 3.4: Sketch discriminator at 64×64 resolution scale: given a sketch attribute vector, the discriminator is trained using the triplets: (i) real-sketch and real-sketch-attributes, (ii) synthesized-sketch and real-sketch-attribute, (iii) wrong-sketch (real sketch but mismatching attributes) and same real-sketch-attributes. Note that the convolutional layers with dashed line are removed when training at lower-resolution discriminator.

shown in Fig. 3.4. In order to learn the discrimination in both image content and semantics, we adopt the triplet matching training strategy [19, 21, 22, 105]. Specifically, given sketch attributes, the discriminator is trained by using the following triplets: (i) real-sketch and real-sketch-attributes, (ii) synthesized-sketch and real-sketch-attributes, and (iii) wrong-sketch (real sketch but mismatching attributes) and same real-sketch-attributes. As shown in Fig. 3.4, two kinds of errors are used to train the discriminator. They correspond to (i) real/face sketch images, and (ii) sketch images and attributes.

The architecture of the proposed sketch discriminator for 64×64 resolution is shown in Fig. 3.4. This architecture can be easily adapted for other resolution scales by adding/removing appropriate the convolutional layers. As shown in Fig. 3.4, two branches with different losses are used to train the discriminator at a certain resolution scale. One consists of a

series of down-sampling convolutional layers (with filter-size 4, stride 2 and padding size 1) to produce a 4×4 probability map and classify each location as true or false. The other branch first embeds the sketch attributes to a $128 \times 4 \times 4$ feature map and concatenates it with the feature maps from the first branch. Another two 1×1 convolutional layers are used to fuse the concatenated feature maps to produce 4×4 probability maps for classification. This branch aims to distinguish whether the semantics in sketch images match the sketch attribute or not, through the feedback loss from another 4×4 probability map.

The overall adversarial loss used to train the network is defined as follows:

$$\begin{aligned}
\mathcal{L}_{s_{Dis}} &= \min_{G_s} \max_{D_s} V(G_s, D_s, \mathbf{X}_s, \mathbf{y}_s, \mathbf{z}_s) \\
&= \sum_{i=1}^{m} \min_{G_s} \max_{D_s^i} (\mathcal{L}^i_{s_{real}} + \mathcal{L}^i_{s_{fake}} + \mathcal{L}^i_{s_{wrong}}), \\
\mathcal{L}^i_{s_{real}} &= \mathbb{E}_{\mathbf{x}_s^i \sim P_{data}(\mathbf{x}_s^i)}[\log D_s^i(\mathbf{x}_s^i)], \\
&\quad + \mathbb{E}_{\mathbf{x}_s^i \sim P_{data}(\mathbf{x}_s^i),, \mathbf{y}_s \sim P_{data}(\mathbf{y}_s)}[\log D_s^i(\mathbf{x}_s^i, \mathbf{y}_s)], \\
\mathcal{L}^i_{s_{wrong}} &= \mathbb{E}_{\mathbf{x}_s^{i'} \sim P_{data}(\mathbf{x}_s^i), \mathbf{y}_s \sim P_{data}(\mathbf{y}_s)}[\log(1 - D_s^i(\mathbf{x}_s^{i'}, \mathbf{y}_s))], \\
\mathcal{L}^i_{s_{fake}} &= \mathbb{E}_{\hat{\mathbf{x}}_s^i \sim P_{G_s(\mathbf{y}_s, \mathbf{z}_s)}}[\log(1 - D_s^i(\hat{\mathbf{x}}_s^i))] \\
&\quad + \mathbb{E}_{\hat{\mathbf{x}}_s^i \sim P_{G_s(\mathbf{y}_s, \mathbf{z}_s)}, \mathbf{y}_s \sim P_{data}(\mathbf{y}_s)}[\log(1 - D_s^i(\hat{\mathbf{x}}_s^i, \mathbf{y}_s))],
\end{aligned} \tag{3.3}$$

where $\hat{\mathbf{x}}_s^i \sim P_{G_s(\mathbf{y}_s, \mathbf{z}_s)}$ stands for the synthesized (fake) sketch image sampled from the sketch generator at scale i, $\mathbf{x}_s^i \sim P_{data}(\mathbf{x}_s^i)$ stands for the real sketch image sampled from the sketch image data distribution at scale i, $\hat{\mathbf{x}}_s^{i'}$ is the attribute-mismatching sketch image sample at scale i, and $\mathbf{y}_s$ is the sketch attribute vector. The total objective loss function is

given as follows

$$\mathcal{L}_{s_{total}} = \sum_{i=1}^{m} \min_{G_s} \max_{D_s^i} (\mathcal{L}^i_{s_{real}} + \mathcal{L}^i_{s_{fake}} + \mathcal{L}^i_{s_{wrong}}) + \lambda_s \mathcal{L}_{s_{aug}}, \tag{3.4}$$

where the hyperparameter λ_s is set equal to 0.01 in our experiments, $\mathcal{L}_{s_{aug}}$ is the KL-divergence regularization in the AA module with sketch attribute $\mathbf{y}_s$ and noise $\mathbf{z}_s$ as inputs.

Stage 2: Sketch-to-Face

Given the synthesized sketches $\hat{\mathbf{X}}_s$ and the facial attributes $\mathbf{y}_f$, the face generator network G_f aims to produce multi-scale outputs as follows

$$G_f(\hat{\mathbf{X}}_s; \mathbf{z}, \mathbf{y}_f) = \{\hat{\mathbf{x}}_f^1, \hat{\mathbf{x}}_f^2, \cdots \hat{\mathbf{x}}_f^m\} \triangleq \hat{\mathbf{X}}_f, \tag{3.5}$$

where $\mathbf{z}$ is noise sampled from a normal Gaussian distribution and $\hat{\mathbf{X}}_f$ are the synthesized facial images with gradually growing resolutions. Similar to the sketch generation network, a set of distinct discriminators are designed for each scale. The overall objective is given as follows:

$$G_f^\star, D_f^\star = arg \min_{G_f} \max_{D_f} V(G_f, D_f, \mathbf{X}_f; \hat{\mathbf{X}}_s, \mathbf{y}_f, \mathbf{z}), \tag{3.6}$$

where $D_f = \{\mathbb{D}_1, \cdots, \mathbb{D}_m\}$ and $\mathbf{X}_f = \{\mathbf{x}_f^1, \cdots, \mathbf{x}_f^m\}$ denote real training images at multiple scales $1, \cdots, m$. In order to preserve the geometric structure of the synthesized sketch from the attribute-to-sketch stage, we adopt the skip-connection architecture from UNet

related works [29, 105, 106]. By using skip-connections, the feature maps from the encoding network are concatenated with the feature maps in the decoding network. This way, the geometric structure of the learned sketch image is inherited in the synthesized facial image. The proposed method is trained end-to-end. The lower-resolution outputs fully utilize the top-down knowledge from the discriminators at higher resolutions. Therefore, the synthesized images from different resolutions preserve the geometric structure, which improves the training stability and synthesis quality.

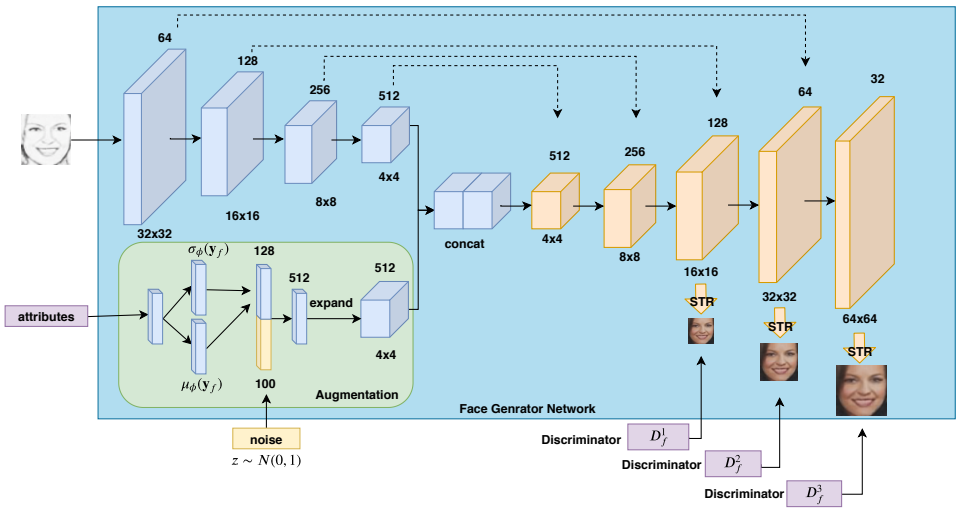


Figure 3.5: The architecture of Face Generator Network. The facial attributes are first embedded by the attribute augmentation module, similar to the one used in stage 1. The synthesized sketch image is also embedded by a sequence of down-sample convolutional layers. These two feature maps are then fused by concatenation. Finally, the fused feature maps are used by the up-sample module to synthesize multi-scale face images.

The architecture of the face generator network is shown in Fig. 3.5. The generator consists of four modules: the AA module, the down-sample module (DO), the UP module, and the STR module. As before, the STR aims convert the feature map into a 3-channel

output image. It consists of two 1×1 convolutional (conv) layers with one Tanh layer. The UP module consists of an up-sampling layer followed by conv-BN-ReLU layers and an additional residual block [102, 103] is between each UP module. The DO module consists of a series of conv-BN-ReLU layers. The overall face generator network architecture consists of the following components DO(64)-DO(128)-DO(256)-DO(512)-AA(512)-UP(512)-UP(256)-UP(128)-UP(64)-UP(32), where the number in round brackets indicate the output channel of feature maps. As shown in Fig. 3.5, the three stretching (STR) modules convert the feature maps into 3-channel output face images at different resolutions.

Discriminator and Training Loss: In the sketch-to-face stage, we use the same architecture of the discriminator as was used in stage 1. We input the triplets as facial images instead of sketch images and replace the sketch attributes by the facial attributes. Furthermore, the training loss function is also the same as the one used in the attribute-to-sketch stage.

Testing

Fig. 5.1 shows the testing phase of the proposed method. A sketch attribute vector $\mathbf{y}_s$ and $\mathbf{z}_s$ sampled from a normal Gaussian distribution are first passed through the sketch generator network G_s to produce a sketch image. Then the synthesized sketch image with the highest resolution, attribute vector $\mathbf{y}_f$ and another noise vector $\mathbf{z}_f$ are passed through the face generator network to synthesize a face image. In other words, our method takes noise and attribute vectors as inputs and generates high-quality face images via sketch images.

3.2 Experimental Results

In this part experimental settings and evaluation of the proposed method are discussed in detail. Results are compared with several related generative models: disCVAE [18], GAN-INT-CLS [19], StackGAN [20], Attribute2Sketch2Face [107], StackGAN++ [21] and HDGAN [22]. The entire network in Fig. 5.1 is trained end-to-end using Pytorch. When training, the learning rate for the generator and the discriminator in the first stage is set equal to 0.0002, while the learning rate in the second stage is set equal to 0.0001.

We conduct experiments using two publicly available datasets: CelebA [30], and deep funneled LFW [108]. The CelebA database contains about 202,599 face images, 10,177 different identities and 40 binary attributes for each face image. The deep funneled LFW database contains about 13,233 images, 5,749 different identities and 40 binary attributes for each face image which are from the LFWA dataset [30].

Note that the training part of our network requires original face images and the corresponding sketch images as well as the corresponding list of visual attributes. The CelebA and the deep funneled LFW datasets consist of both the original images and the corresponding attributes. To generate the missing sketch images in the CelebA and the deep funneled LFW datasets, we use a public pencil-sketch synthesis method [1] to generate the sketch images from the face images. Fig. 3.6 shows some sample generated sketch images from the CelebA and the deep funneled LFW datasets.

The MTCNN method [109] was used to detect and crop faces from the original images.

[1] http://www.askaswiss.com/2016/01/how-to-create-pencil-sketch-opencv-python.html

Figure 3.6: Sketch images sampled from the LFW and the CelebA datasets are shown in row 1 and row 2 respectively.

The detected faces were scaled to the size of 64×64. Since many attributes from the original list of 40 attributes were not significantly informative, we selected 23 most useful attributes for our problem. Furthermore, the selected attributes were further divided into 17 texture and 6 color attributes as shown in Table 3.1. During experiments, the texture attributes were used to train the sketch generator network while all 23 attributes were used to train the face generator network.

Table 3.1: List of fine-grained texture and color attributes.

Texture	5_o_Clock_Shadow, Arched_Eyebrows, Bags_Under_Eyes, Bald, Bangs, Big_Lips, Big_Nose, Bushy_Eyebrows, Chubby, Eyeglasses, Male, Mouth_Slightly_Open, Narrow_Eyes, No_Beard, Oval_Face, Smiling, Young
Color	Black_Hair, Blond_Hair, Brown_Hair, Gray_Hair, Pale_Skin, Rosy_Cheeks

CelebA Dataset Results

The CelebA dataset [30] consists of 162,770 training samples, 19,867 validation samples and 19,962 test samples. We combine the training and validation splits together to

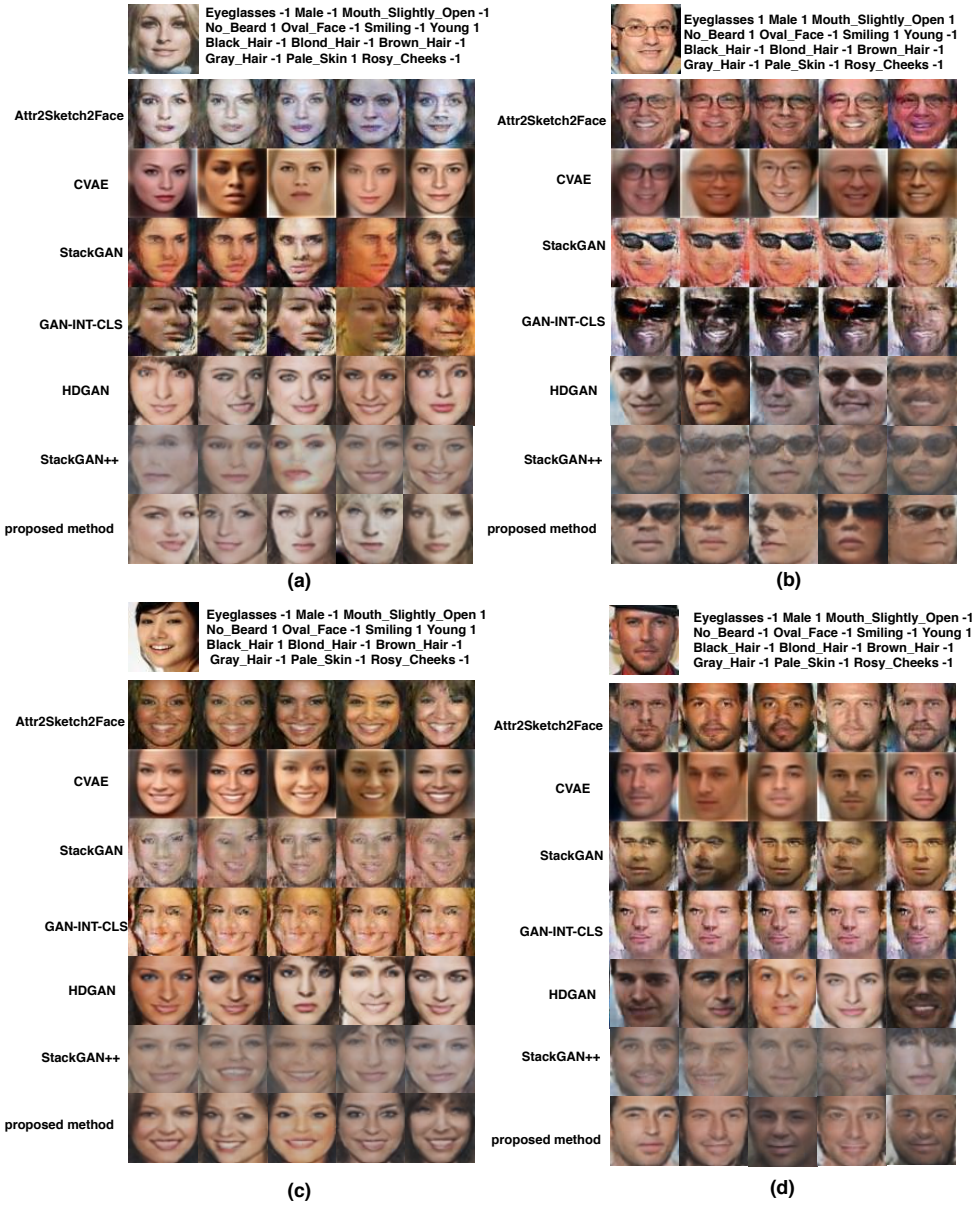


Figure 3.7: Image generation results on the CelebaA dataset. First row of each sub-figure shows the reference image and its corresponding attributes. The images generated by different methods are shown in different rows.

train our models. After detection and alignment, we obtain 182,468 samples which we use for training our proposed model. During training, we use the batch size of 40. The ADAM algorithm [110] with learning rate of 0.0002 is used to train the network. In total, 20 training epoch are used during training and the initial learning rate is frozen for the first 10 epochs. For the next 10 epochs, we let it drop by 0.1 of the initial value after every epoch. The latent feature dimension for sketch/facial attribute is set equal to 128. The noise vector dimension is set equal to 100. Three scales (16×16, 32×32, and 64×64) are used in our multi-scale network.

Sample image generation results corresponding to different methods from the CelebA are shown in Fig. 3.7. For fair comparison with those stage-wise training algorithm, we adopt StackGAN [20] network to two scale resolution 32×32 and 64×64. Moreover, we adopt StackGAN++ [21] HDGAN [22] in the same resolution scales: 16×16, 32×32, and 64×64. Note that these results are obtained by inputting a certain attribute vector along with random noise. As can be seen from this figure, GAN-INT-CLS and StackGAN methods easily meet the modal collapse issue in this problem. During training, the generator learns to generate a limited number (1 or 2) of image samples corresponding to a certain list of attributes. This synthesized results are good enough to fool the discriminator. Thus, the generator and discriminator networks do not get optimized properly. The disCVAE method is able to reconstruct the images without model collapse but they are blurry due to the imperfect L_2 measure in the Gaussian distribution loss. In addition, some of the attributes are difficult to see in the reconstructions corresponding to the disCVAE method,

such as the hair color. This is because of the imperfect latent embedding by the variational bound limitation in VAE. Also, the other Attribute2Sketch2Face is able to generate realistic results, but the image quality is slightly inferior. The recent state-of-art text-to-image synthesis approaches (stackGAN++ and HDGAN) generate plausible facial images from visual attributes. However, the generated facial images do not always preserve the corresponding attributes very well. Compared with all the baselines, the proposed method not only generates realistic facial images but also preserves the attributes better than the others. We believe that this is mainly due to the way we appraoch the attribute-to-face synthesis apporach by decomposing it into two problems, attribute-to-sketch and sketch-to-face. By factoring the original problem into two separate problems, the model at each stage learns better conditional data distribution. Furthermore, the use of multi-scale generators in the proposed GANs also help in improving the performance of our method.

LFW Dataset Results

Images in the LFWA dataset come from the LFW dataset [108], [111], and the corresponding attributes come from [30]. This dataset contains the same 40 binary attributes as in the CelebA dataset. After pre-processing, the training and testing subsets contain 6,263 and 6,880 samples, respectively. We use all the training splits to train our model. The ADAM algorithm [110] with learning rate of 0.0002 is used for both generators and discriminators. The initial learning rate is frozen in the first 100 epochs and is then dropped by 0.01 for the remaining 100 epochs. All the other experimental settings are the same as

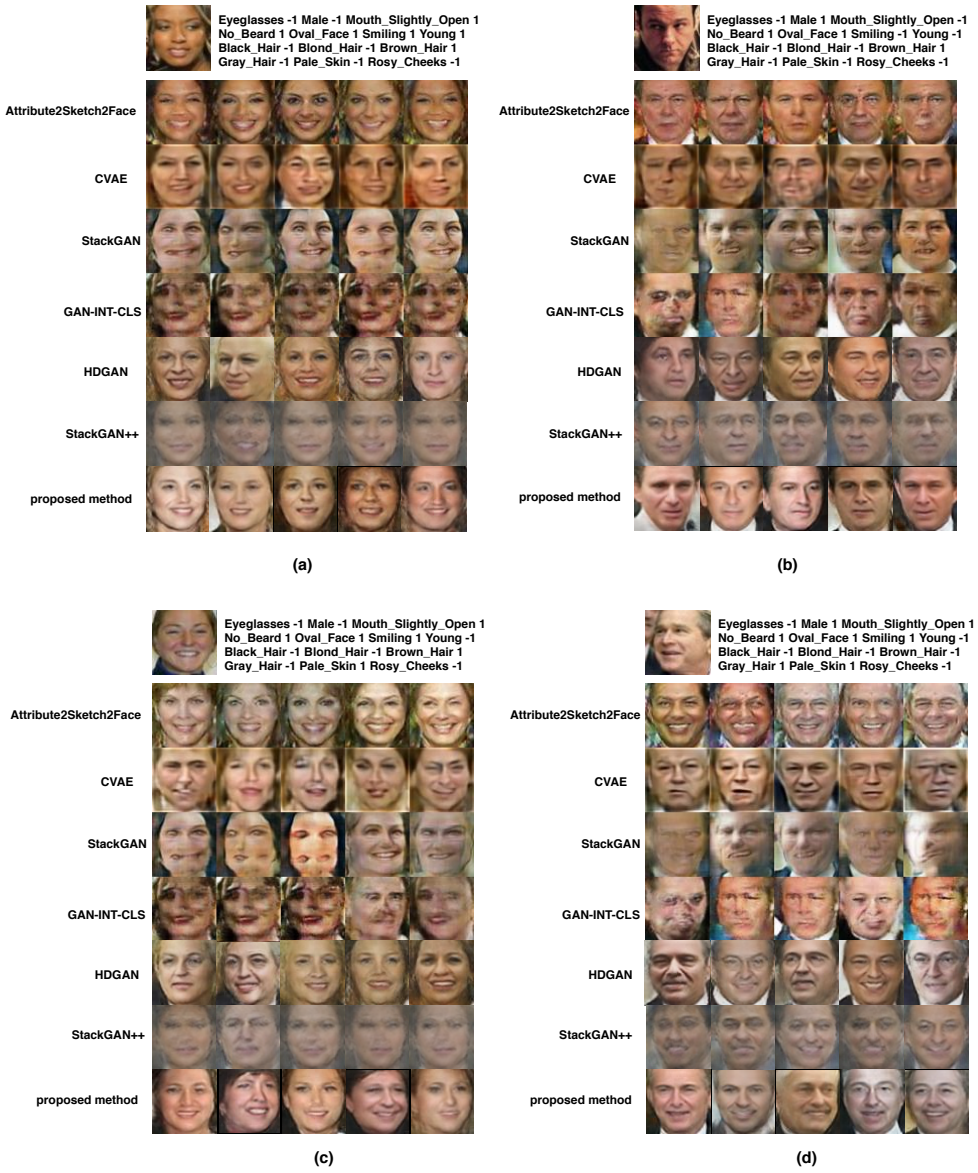


Figure 3.8: Image generation results on the LFWA dataset. First row of each sub-figure shows the reference image and its corresponding attributes. The images generated by different methods are shown in different rows.

the ones used in with CelebA dataset.

Sample results corresponding to different methods on the LFWA dataset are shown in Fig. 3.8. For fair comparison, the multi-scale resolution settings are the same as used with the experiments on the CelebA dataset. In particular, we use 32×32 and 64×64 resolution scales for StackGAN [20] training and 16×16, 32×32 and 64×64 multiple resolution scales for HDGAN [22] and StackGAN++ [21] as well as our proposed method. The disCVAE method produces reconstructions which are blurry. Previous conditional GAN-based approaches such as GAN-INT-CLS [19] and StackGAN [20] also produce poor quality results due to the model collapse during training. Recent StackGAN++ and HDGAN works generate plausible facial images (HDGAN is better at the color diversity). The previous work Attribute2Sketch2Face, which is a combination of CVAE and GAN, is also able to generate facial images with corresponding attributes. However, the proposed method is able to reconstruct high-quality attribute-preserved face images better than the previous approaches.

CelebA-HQ Dataset Results

In order to demonstrate how our proposed method works on high-resolution images, we also conduct an experiment using a recent proposed CelebA-HQ dataset. The CelebA-HQ dataset [101] is a high-quality version of the CelebA dataset, which consists of 30,000 images with 1024×1024 resolution. Due to GPU and memory limitations, we conduct experiments on 256×256 resolution images and compare the performance with Stack-

GAN++ [21] and HDGAN [22]. The reason why we chose these two baselines is due to their capability to deal with high resolution images. Sample results are shown in Figure 3.9.

For fair comparison, we set the number of resolution scale $s = 3$ for all methods. In order to adopt our method to the high-resolution dataset, we follow the strategy that removing/adding the number of UP/DO block (as defined in Section 3.1) in the generator and the discriminator. In particular, we set the STR modules at resolution 64×64, 128×128 and 256×256 respectively. In experiments, the batch-size is set equal to 16 for our proposed method, which is smaller than StackGAN++ and HDGAN, which are set equal to 24, due to the GPU memory limitations. Also, when training on this dataset, we train the sketch generator first and then use the pre-trained model for training the face generator.

As can be seen from Figure 3.9, our proposed method can synthesize photo-realistic images on high-resolution images as well. Moreover, when we compare the attributes from the synthesized images with the given attributes, we can observe that our method preserves the attributes better than the other methods. Quantitative comparisons in terms of the FID scores also show that the proposed method performs favorably compared to StackGAN++ and HDGAN. In addition, comarison of our method in Table 3.2 with only a single scale shows the significance of our multi-scale network.

Face Synthesis

In this part, we show the image synthesis capability of our network by manipulating the input attribute and noise vectors. Note that, the testing phase of our network takes attribute

Methods	FID score
HDGAN [22]	114.912
StackGAN++ [21]	35.988
Single-scale (proposed method)	37.381
Proposed method	30.566

Table 3.2: Quantitative results (FID scores) corresponding to different methods on the CelebA-HQ dataset.

StackGAN++ [21] HDGAN [22] Proposed method

Figure 3.9: Image synthesis results on 256×256 resolution. The attributes used to generate these images are: Eyeglasses -1 Male -1 Mouth_Slightly_Open -1 No_Beard 1 Oval_Face -1 Smiling -1 Young 1 Black_Hair -1 Blond_Hair -1 Brown_Hair -1 Gray_Hair -1 Pale_Skin 1 Rosy_Cheeks -1.

Table 3.3: Quantitative results corresponding to different methods.The FID score and Attribute L_2 measure are used to compare the performance of different methods.

	LFW		CelebA	
Baselines	FID Score	Attribute L_2	FID Score	Attribute L_2
GAN-INT-CLS [19]	85.811	0.093 ± 0.027	92.793	0.104 ± 0.024
disCVAE [18]	103.855	0.086 ± 0.019	91.012	0.080 ± 0.042
StackGAN [20]	70.379	0.085 ± 0.029	63.816	0.091 ± 0.021
StackGAN++ [21]	50.360	0.059 ± 0.026	49.889	0.061 ± 0.026
HDGAN [22]	48.930	0.053 ± 0.023	43.206	0.056 ± 0.020
Attribute2Sketch2Face [107]	60.487	0.059 ± 0.034	58.896	0.067 ± 0.022
Attribute2Sketch2Face-v2 (proposed method)	**43.712**	$\mathbf{0.048 \pm 0.020}$	**33.497**	$\mathbf{0.051 \pm 0.019}$

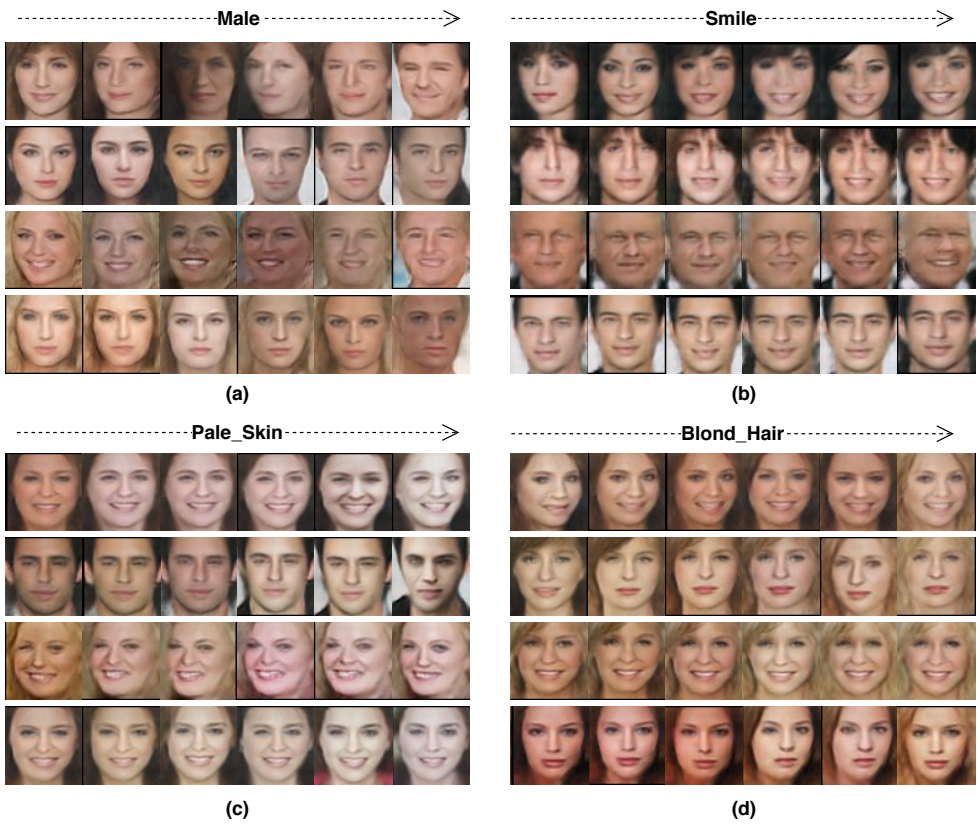


Figure 3.10: Facial image progressive synthesis on CelebA when attributes are changed. These progressive changes are based on one certain attribute manipulating while the others are keep frozen. (a) Male. (b) Smile. (c) Original skin tone to pale skin tone. (d) Original hair color to black hair color.

Figure 3.11: Facial image synthesis sampled when attributes are kept frozen while the noise vector is changed. Note that the identity, pose, or facial shape changes as we vary the noise vector but he attributes stay the same on the synthesized images.

vector and noise as inputs and produces synthesized face as the output. In the first set of experiments with image synthesis, we keep the random noise vector frozen and change the weight of a particular attribute as follows: $[-1, -0.1, 0.1, 0.4, 0.7, 1]$. The corresponding results on the CelebA dataset are shown in Fig. 3.10. From this figure, we can see that when we give higher weights to a certain attribute, the corresponding appearance changes. For example, one can synthesize an image with a different gender by changing the weights corresponding to the gender attribute as shown in Fig. 3.10(a). Each row shows the progression of gender change as the attribute weights are changed from -1 to 1 as described above. Similarly, figures (b), (c) and (d) show the synthesis results when a neutral face image is transformed into a smily face image, skin tones are changed to pale skin tone, and hair colors are changed to black, respectively. It is interesting to see that when the attribute weights other than the gender attribute are changed, the identity of the person does not change too much.

In the second set of experiments, we keep the input attribute vector frozen but now change the noise vector by inputing different realizations of the standard Gaussian. Sample results corresponding to this experiment are shown in Fig. 3.11 using the CelebA. Each column shows how the output changes as we change the noise vector. Different subjects are shown in different rows. It is interesting to note that, as we change the noise vector, attributes stay the same while the identity changes. This can be clearly seen by comparing the synthesized results in each row.

Quantitative Results

In addition to the qualitative results presented in Fig. 3.7, 3.8, we present quantitative comparisons in Table 3.3. Since the ground-truth images corresponding to the noise-generated images are not available, we choose the quantitative criterion based on the Frchet Inception Distance (FID) [112, 113] and Attribute L_2-norm. The FID is a measure of similarity between two datasets of images. It was shown to correlate well with human judgment of visual quality and is most often used to evaluate the quality of samples generated by GANs. Attribute L_2-norm is used to compare the quality of attributes corresponding to different images. We extract the attributes from the synthesized images as well as the reference image using the MOON attribute prediction method [93]. Once the attributes are extracted, we simply take the L_2-norm of the difference between the attributes as follows

$$\text{Attribute } L_2 = \|\hat{a}_{ref} - \hat{a}_{synth}\|_2, \tag{3.7}$$

where $\hat{a}_{ref}$ and $\hat{a}_{synth}$ are the 23 extracted attributes from the reference image and the synthesized image, respectively. Note that lower values of the FID score and the Attribute L_2 measure imply the better performance. The quantitative results corresponding to different methods on the CalebA and LFW datasets are shown in Table 3.3. Results are evaluated on the test splits of the corresponding dataset and the average performance along with the standard deviation are reported in Table 3.3.

As can be seen from this table, the proposed method produces the lowest FID scores

implying that the images generated by our method are more realistic than the ones generated by other methods. Furthermore, our method produces the lowest Attribute L_2 scores. This implies that our method is able to generate attribute-preserved images better than the other compared methods. This can be clearly seen by comparing the images synthesized by different methods in Fig. 3.7 and Fig. 3.8.

3.3 Summary

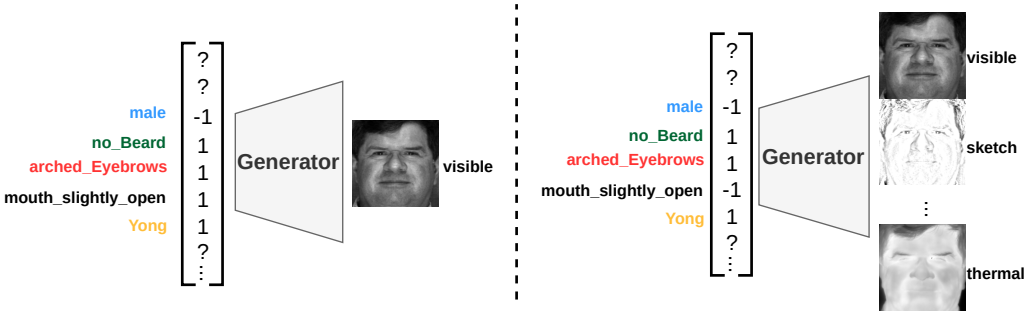
We presented a novel deep generative framework for reconstructing face images from visual attributes. Our method makes use of an intermediate representation to generate photo realistic images. The training part of our method consists of two models: Sketch Generator Network and Face Generator Network. Multi-scale hierarchical network architectures are proposed for each generator networks. Various experiments on three publicly available datasets show the significance of the proposed synthesis framework. In addition, an ablation study was conducted to show the importance of different components of our network. Various experiments showed that the proposed method is able to generate high-quality images and achieves significant improvements over the state-of-the-art methods.

Chapter 4

Multimodal Face Synthesis from Visual Attributes

As discussed earlier, generating face images from visual attributes is an important problem in the biometrics and computer vision communities due to its applications in forensics, entertainment, and law enforcement (see Fig. 4.1(a)). Recent advances in deep generative networks have made it possible to generate high-quality face images from facial attributes [18,21,22,114–116]. However, existing attribute-to-face synthesis methods mainly focus on generating unimodal face images (i.e visible faces) from attributes. In many scenarios, the gallery images contain multiple modalities and the domain gap between the generated unimodal face images and gallery images will degrade the recognition performance. Therefore, a multimodal attribute-to-face synthesis method can assist law enforcement officers to identify a person regardless the domain gap by simultaneously generating face

images in visible, sketch and thermal domains.



(a) Unimodal synthesis (b) Multimodal synthesis

Figure 4.1: Illustration of unimodal and multimodal face generation from visual attributes. Given a list of facial attributes, we aim to use a single generator to simultaneously synthesize multimodal face images with consistant geometry and texture patterns such that they satisfy these attributes.

In this work, we focus on the problem of simultaneously synthesizing multimodal face images from visual attributes (see Fig. 4.1(b)). A naive solution to this problem is to simply use an attribute-to-face synthesis method [18, 21, 22, 114, 115] to generate a visible image from facial attributes and then use an image-to-image translation method [117–121] to synthesize images from the visible domain to the other domains such as thermal or composite sketch. However, it is well-known in the biometrics community that synthesizing facial images from one domain to another (i.e. visible to thermal) itself is an extremely difficult problem and often leads to poor synthesis results [14, 34, 35, 61–63, 82, 122–130]. Hence, a combination of attribute-to-face synthesis with cross-modal synthesis will not be an effective approach for this problem.

Another approach would be to train c unimodal attribute-to-face synthesis methods separately for c different modalities. However, in this case one often looses geometric and texture consistency among the multimodal face images when the generators are trained

separately for each modality. Hence, the synthesized multimodal face images do not contain consistant identities. Furthermore, training c different networks corresponding to c modalities will require large memory and computation time.

We propose a new generative adversarial network (GAN), called Att2MFace, that can directly synthesize multimodal face images from visual attributes. The generator network consists of multimodal stretch-out modules which convert the modality invariant features to multimodal images. On the other hand, the discriminator network contains stretch-in modules which convert the multimodal images to a modality invariant feature representation which can be used to discriminates between real and fake images. In addition, an auxiliary estimator is used along with the discriminator to estimate the probability of the target attributes. By back-propagating the errors of image discriminability and attribute probability, the generator can learn to synthesize a diverse set of realistic multimodal face images from visual attributes.

One of the main advantages of the proposed Att2MFace model is that it can generate multimodal images even if there are no paired multimodal images available for training. This is due to the following two reasons: (1) multimodal images are generated from a common feature representation by the stretch-out module, and (2) the texture pattern in lower resolution is inherited to a higher resolution by the progressive-growth training. This unsupervised learning setting is convenient for many applications because collecting paired multimodal face images is laborious and expensive. Furthermore, this unsupervised learning shows its benefits in improving the diversity of generated images [23]. For example,

given unpaired visible and sketch images, the generator can explore the unseen texture pattern (complementary information) in the sketch domain while learning to synthesize in the visible domain.

4.1 Proposed Method

In this chapter, we provide details of the proposed Att2Mface modal, which consists of a multimodal generator and a multimodal discriminator with an auxiliary attribute estimator. Given a visual attribute vector $\mathbf{y}_a$ and a noise vector $\mathbf{z}$ sampled from the normal distribution, the proposed generator G aims to simultaneously generate multimodal images $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \cdots, \hat{\mathbf{x}}_c$ as follows

$$G(\mathbf{z}, \mathbf{y}_a) = \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \cdots, \hat{\mathbf{x}}_i, \cdots, \hat{\mathbf{x}}_c\}, \tag{4.1}$$

where c indicates the total number of face modalities. Fig. 4.2 gives an overview of the proposed network in details. In what follows, we provide details of the different modules in our framework.

4.1.1 MultiModal Generator

The proposed generator network consists of the following components: multi-layer perceptron (MLP), Initial Block, Transition Blocks and Multi-modal Stretch-out modules. The MLP module maps $\mathbf{y}_a$ and $\mathbf{z}$ to an attribute latent space. MLP consists of one linear fully-connected layer followed by a LeakyReLU function. The Initial Block learns the feature

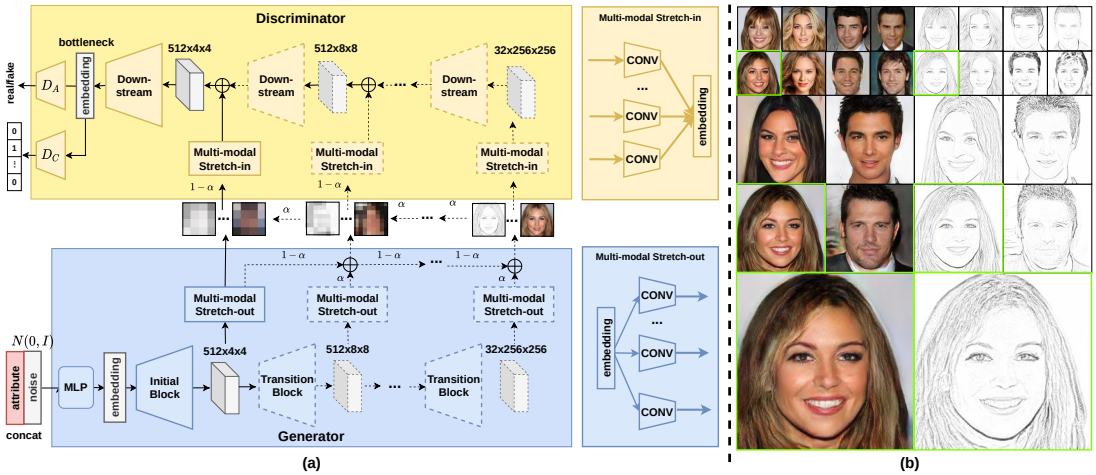


Figure 4.2: An overview of the proposed Att2MFace framework. (a) Details of the generator and discriminator networks with different modules. The progressive training parts are indicated by dashed lines. (b) Synthesized multimodal images corresponding to the same input at different resolution scales based on progressive training are shown in green boxes.

maps in a particular resolution (i.e. 4×4) from the attribute features. The Initial Block consists of one reshape function and one convolutional layer followed by a LeakyReLU function. Then, another Transition Block is implemented to map the feature maps onto a higher resolution (i.e. 8×8) by utilizing an upsampling layer and two convolutional layers. In addition, pixel-wise feature equalization [38, 131] is employed after all the convolutional layers in the generator. Pixel-wise feature equalization is defined as follows

$$\mathcal{F}_{m,n} = \tilde{\mathcal{F}}_{m,n} / \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\tilde{\mathcal{F}}^{i}_{m,n})^2 + \epsilon}, \tag{4.2}$$

where $\epsilon = 10^{-8}$, N is the number of feature maps and $\mathcal{F}_{m,n}$ and $\tilde{\mathcal{F}}_{m,n}$ are the normalized and original feature vectors in pixel (m, n), respectively. Given a common feature map output from the transition block at a certain resolution scale, a multimodal stretch-out module that contains c convolution operations are implemented to simultaneously generate face images in different modalities. By leveraging the same modality invariant content from the

common feature map, this module aims to add modality specific style when converting the feature maps to a output images.

4.1.2 MultiModal Discriminator

As shown in Fig. 4.2, the multimodal discriminator, D contains two output streams: estimation D_C and authentication D_A. The authentication stream, D_A, measures the probability of whether a given multimodal sample belongs to the data distribution. On the other hand, the estimation stream, D_C, aims to learn the mapping from the input image to the corresponding target label probability distribution, which is achieved by adding an auxiliary estimator. This estimator allows a single discriminator to learn the discriminability based on the corresponding multimodal image labels [132].

The discriminator contains a series of Multimodal Stretch-in and Down-stream modules. For each resolution scale, a multimodal stretch-in module is employed which contains c convolution operators followed by LeakyReLU. This module distills the modality specific style and converts the three-channel input to modality invariant multi-channel representation. Given the feature maps from the multi-modal stretch-in module, the Down-stream module learns a representation at different resolutions. The Down-stream module consists of one average pooling layer and two convolution layers followed by feature equalization. Finally, after a series of Down-stream blocks, the bottleneck feature maps are of size 4×4 and are passed on to both the estimation stream, D_C and the authentication steam, D_A. The authentication and estimation streams contain two fully-connected layers to learn the

probability of discrimination and the corresponding target label, respectively.

To make sure that the estimator learns both the modality and attribute probabilities, we define a new target label $\mathbf{y}$ as follows:

$$\mathbf{y}_i = [\mathbf{y}_a, \mathbf{y}_m], \quad \text{where} \quad \mathbf{y}_m = [0, 1, \cdots, 0], \tag{4.3}$$

where, $\mathbf{y}_a$ is the visual attribute vector, $\mathbf{y}_m = [0, 1, \cdots, 0]$ is a c-dimensional one-hot vector indicating what modality the image belongs to, and c is the total number of modalities.

4.1.3 Progressive Training

Recent image generation approaches have found progressive training beneficial [38, 133, 134]. In progressive training, the idea is to start with low-resolution images and then progressively increase the resolution by adding layers to the network. This way, large-scale structure of the image distribution is first discovered and then finer details are added by additional layers. Hence, we start training our model from resolution 4×4 and progressively grow it to $8 \times 8, \cdots, 256 \times 256$. To achieve this, skip connection between the output of newly added Transition/Down-stream Block and the existing output/input of the generator/discriminator are balanced by a trainable weight parameter α. Starting from a light weight α ($\alpha = 0$) helps to smooth out the influence of adding a new Transition/Down-stream block to the network. Additionally, the equalization operations as defined in Eq. 4.2 are used in both generator and discriminator networks to prevent the escalation of feature magnitudes.

4.1.4 Objective Function

The objective functions used to train the multimodal generator and the discriminator networks consist of the adversarial loss and the classification loss for authentication and estimation streams, respectively.

Adversarial Loss: In order to make the generated multimodal images indistinguishable from real multimodal images, we adopt the WGAN-GP loss [135, 136] which is defined as follows:

$$\mathcal{L}_{adv} = \sum_{i=1}^{c} (\mathbb{E}[D_A(\mathbf{x}_i)] - \mathbb{E}[(D_A(\hat{\mathbf{x}}_i))] - \lambda_{gp}\mathbb{E}[(\|\nabla_{\mathbf{x}^\star} D_A(\mathbf{x}_i^\star)\|_2 - 1)^2]), \tag{4.4}$$

where $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ correspond to the real image and the synthesized image corresponding to the i-th modality, respectively. Here, $\mathbf{x}_i^\star$ is sampled uniformly along a straight line between a pair of real $\mathbf{x}_i$ and the generated $\hat{\mathbf{x}}_i$ [137]. D_A refers to the output probability score from the authentication stream. The discriminator attempts to maximize this objective, while the generator attempts to minimize it. We set $\lambda_{gp} = 10$ in our experiments.

Classification Loss: The classifier D_C aims to learn the mapping from the input images to the target label probability distribution. On the other hand, the generator G aims to synthesize images that match the corresponding target label. In order to achieve this, the classification loss is imposed when optimizing both D_C and G. The objective consists of two separate terms: (i) the classifier D_C is optimized by the real images $\mathbf{x}_i$ and the target labels $\mathbf{y}_i$, and (ii) the generator G is optimized to synthesize $\hat{\mathbf{x}}_i$ with corresponding target

label $\mathbf{y}_i$. In particular, the first term is defined as follows:

$$\mathcal{L}_{cls}^{real} = \sum_{i=1}^{c} \mathbb{E}_{\mathbf{x}_i,\mathbf{y}_i}[-\log P(D_C(\mathbf{x}_i) = \mathbf{y}_i|\mathbf{x}_i)], \tag{4.5}$$

where $D_C(\mathbf{x}_i)$ denotes the class probability vector over the classification module D_C given real input $\mathbf{x}_i$. By minimizing this objective with the target label $\mathbf{y}_i$, the discriminator learns to classify a real image to its correct target label. For second term, the loss function is defined as follows

$$\mathcal{L}_{cls}^{fake} = \sum_{i=1}^{c} \mathbb{E}_{\hat{\mathbf{x}}_i,\mathbf{y}_i}[-\log P(D_C(\hat{\mathbf{x}}_i) = \mathbf{y}_i|\hat{\mathbf{x}}_i)], \tag{4.6}$$

where $\hat{\mathbf{x}}_i \sim G(\mathbf{z}, \mathbf{y}_a)_i$ is the i-th modality fake image. The generator G attempts to minimize this classification objective so that the synthesized image has the corresponding target label $\mathbf{y}_i$.

Overall Objective: The overall objective function to optimize the discriminator D and the generator G is defined as follows

$$\begin{aligned} \mathcal{L}_D &= -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls}^{real} \\ \mathcal{L}_G &= \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls}^{fake}, \end{aligned} \tag{4.7}$$

where λ_{cls} is the parameter that controls the contribution of the classification losses. In our experiments, we have found that a larger value of λ_{cls} leads to a slower training convergence and a smaller value of λ_{cls} does not make the synthesized images preserve the corresponding attributes well. We set $\lambda_{cls} = 1$ in our experiments as it gives us better performance.

During training, the generator and discriminator networks are optimized iteratively. When updating the discriminator, D is optimized by maximizing the difference from synthetic data to real data distribution (authentication D_A). On the other hand, it is also op-

timized by estimating the corresponding ground-truth target label (estimation D_C). When updating the generator, G is optimized by minimizing the difference from the synthetic data to real data distribution. It is also optimized by synthesizing data samples that match the corresponding target label. In this way, the synthesized images are not only photo-realistic but also satisfy the correct target label.

4.1.5 Network Architecture

The architectures corresponding to the multimodal generator and multimodal discriminator are shown in Table 4.1 and Table 4.2. The values the noise code $\mathbf{z} \in \mathcal{R}^{512}$ and the attribute vector $\mathbf{y}_a \in \mathcal{R}^{d_a}$ take are in the range $[-1, 1]$. Both networks consist of various Transition and Down-stream Blocks. Both the Multimodal Stretch-out and Stretch-in modules consist of c numbers of 1×1 convolutional layers corresponding to c different modalities. Except for the stretch-out/stretch-in module on the 4×4 resolution, the multimodal output images are linearly interpolated with the outputs from the Conv 1×1 layer with the upsampled/downsampled ones from the former resolution scale by the trainable weight α.

We use the Leaky ReLU operation with leakiness 0.2 in all the layers of both generator and discriminator. Besides, we use pixel-wise equalization instead of batch normalization or instance normalization after each Conv 3×3 layer in both networks. Upsample and Downsample operations utilize 2×2 element replication and average pooling, respectively. At the end of the discriminator, two fully-connected layers are utilized for authentication

D_A and classification D_C individually. The output neurons of these two fully-connected layers are trained with the WGAN-GP and cross-entropy loss separately.

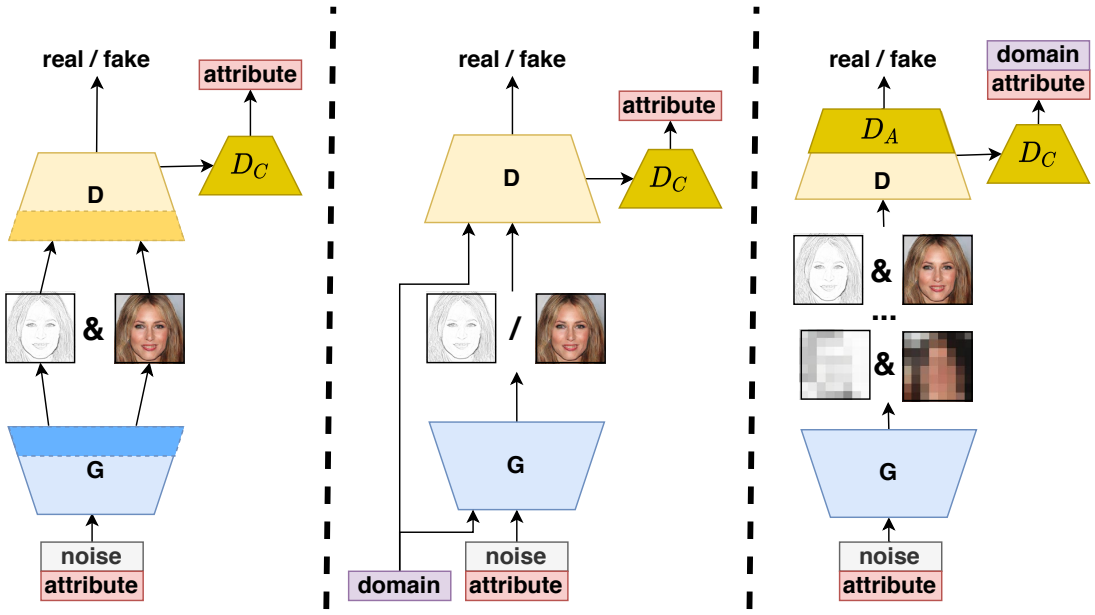
For the other baseline networks, we use the noise code and the attribute code with the same dimension. Additionally, we follow the training configuration described in those respective papers as closely as possible. For CoGAN, we extend the number of output branches to c modalities by replication. The classifier is adopted with the cross-entropy loss. Except for the fist layer of the discriminator and the last layer of the generator, the rest of layers are all with the tied-weights. For RegCGAN, we extend the domains from 2 in the original paper to c. We obtain the multimodal face images during testing by changing the domain label while keeping the noise and attribute code the same.

4.2 Experiments

In this part, the experimental evaluations of the proposed method are discussed in detail. We compare Att2MFace against several multimodal synthesis baselines both qualitatively and quantitatively. In addition, we present attribute and noise manipulation results.

4.2.1 Baseline Models

We consider the following two recently proposed multimodal image synthesis methods as baselines: CoGAN [24] and RegCGAN [138]. Both of these methods perform noise-to-image synthesis over multiple modalities. We concatenate visual attributes with noise



(a) CoGAN [24] (b) RegCGAN [138] (c) Att2MFace

Figure 4.3: Comparison among baseline models.

to generate attribute conditioned multimodal images. In another baseline, we use an output from an attribute-to-face synthesis network as an input to multimodal image-to-image translation network such as StarGANv2 [139]. Fig. 4.3 presents a comparison among these baseline methods. In what follows, we give more details regarding these baselines.

CoGAN [24] utilizes coupled generators/ discriminators that share weights in shallow/deeper layers to synthesize multimodal images. The generators and discriminators are jointly optimized. We adopt the architecture of CoGAN with a classifier to predict the attribute. Except for the last layer of the generator and the first layer of the discriminator, the rest of the layers from each modality network are tied.

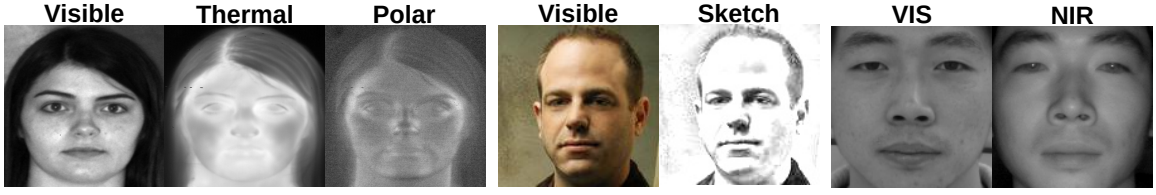
RegCGAN [138] generates multimodal images by changing the modality label $\mathbf{y}_m$. Training this model is regularized by forcing the first layer generator features to be similar for

paired inputs and forcing the last layer features of the discriminator to be similar for paired inputs. Similar to CoGAN, an auxiliary classifier is attached to the discriminator for predicting the attributes.

StarGANv2* [139] is the state-of-the-art multimodal image-to-image translation model. We re-trained StarGANv2 based on different modalities available in each dataset. For a fair comparison, visible images from our Att2MFace network are used as input to StarGANv2. This network synthesizes the remaining multimodal images from the given visible image.

4.2.2 Datasets

The following three multimodal face datasets are used to conduct experiments: ARL Multi-modal Face Dataset [34, 140], CelebA-HQ dataset [38] and CASIA NIR-VIS 2.0 Face Database [39]. Sample images from different modalities from these datasets are shown in Fig. 4.4. The list of facial attributes used from these datasets to synthesize multimodal face images is tabulated in Table 4.4. During training, multimodal images are randomly sampled according to the attributes. Hence, multimodal data are not required in pairs.



(a) (b) (c)

Figure 4.4: Sample images and the corresponding modalities from (a) ARL Multi-modal Face Dataset [34, 140], (b) CelebA-HQ [38], and (c) CASIA NIR-VIS 2.0 [39].

ARL Multimodal Face Dataset: The ARL multimodal dataset [34, 140] consists of 5,419 polarimetric thermal and visible pairs of images from 121 subjects in various expressions, pose, etc. conditions. We resize the images to 256×256 resolution for training our model.

CelebA-HQ: The CelebA-HQ dataset [38] consists of 30,000 high-resolution face images (i.e. 1024× 1024) with corresponding 40 facial attributes [30]. We extract sketch images [2] from the visible images and view them as the second modality images. We resize the images to 256×256 resolution for training our model.

CASIA NIR-VIS 2.0: The CASIA NIR-VIS 2.0 Face Dataset [39] contains 12,487 near-infrared (NIR) images and 5,093 visible images corresponding to 725 subjects. The images in this dataset have been tightly cropped with 128×128 resolution.

4.2.3 Implementation

The Adam optimizer [110] with a batch size of 16 is used to train the network. The learning rate starts from 0.001 for the generator and the discriminator. The number of iteration for seven scales $4 \times 4, 8 \times 8, \cdots, 256 \times 256$ are set equal to $4.8 \times 10^4, 9.6 \times 10^4, \cdots 9.6 \times 10^4, 2.0 \times 10^5$, respectively. It takes around 8 days to train the entire network on two NVIDIA TITAN Xp GPUs. The hyperparameters are selected based on the lowest FID score from 5000 random samples during training.

In CoGAN [24], the multimodal face images are synthesized by concatenating the noise vector with visual attributes. In RegCGAN [138], different modality images are obtained

[2] http://www.askaswiss.com/2016/01/how-to-create-pencil-sketch-opencv-python.html

by feeding noise and attribute concatenated vector with a specific modality label c different times. Regarding StarGANv2* [139], the model is first re-trained based on particular modalities and then the synthesized visible image from Att2MFace is used as input to generate multimodal images corresponding to other modalities.

4.2.4 Results

We compare our method with the baseline models qualitatively and quantitatively. The performance of different methods is quantitatively evaluated using the Fréchet inception distance (FID) [112] and Learned Perceptual Image Patch Similarity (LPIPS) distance [141]. A lower FID value implies that the generated data are closer to the real data. When computing the FID score, we choose the same number of synthetic samples as the number of original images in the related dataset. Higher LPIPS values imply that the synthesized images are more diverse. Following [117, 142, 143], we randomly choose 2k synthetic images and measure the pairwise LPIPS for each modality.

Image Quality: The quantitative results corresponding to different methods are shown in Table 4.3. Figs. 4.5, 4.6, & 4.7 show the qualitative performance of different methods on the ARL Multi-modal Face Dataset, CelebA-HQ and CASIA NIR-VIS 2.0 datasets, respectively. As shown in these figures, Att2MFace is able to synthesize multimodal images directly from visual attributes better than the other methods. In particular, the generated images preserve the attributes that were used to synthesize images (listed on the first column). Regarding the other baseline methods, CoGAN [24] and RegCGAN [138]

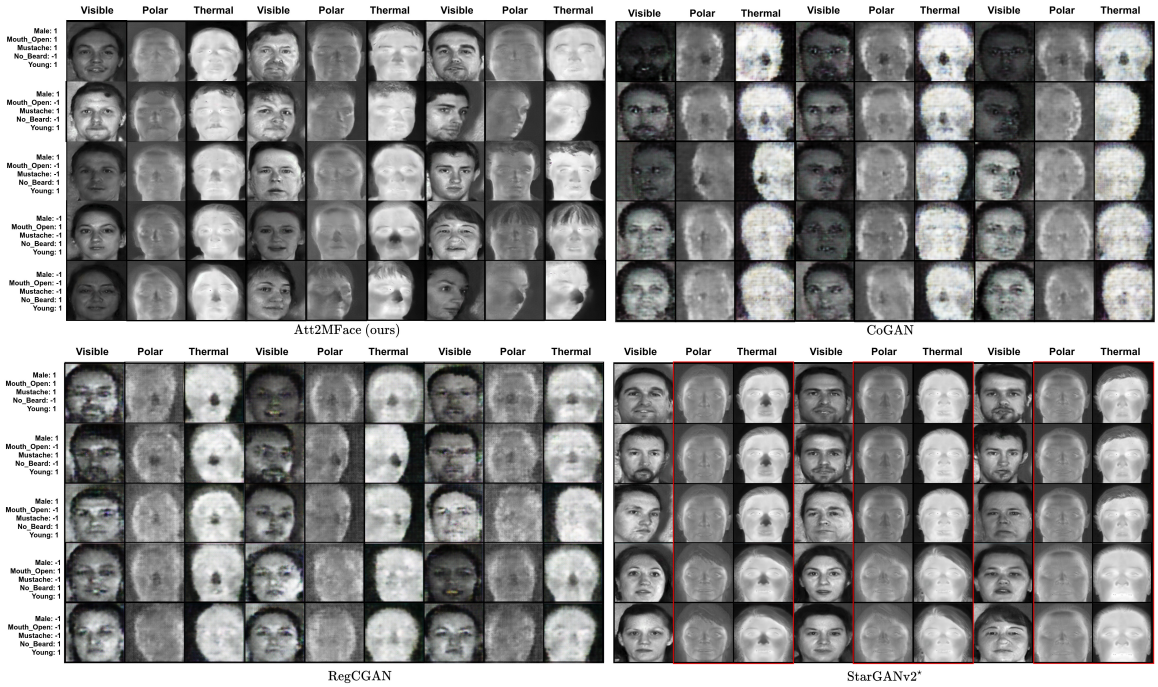


Figure 4.5: Sample 256 × 256 resolution multimodal images generated by different methods using the ARL Multimodal Face Database (Zoom-in for better visualization). The red boxes highlight the loss of geometry consistency by StarGANv2.

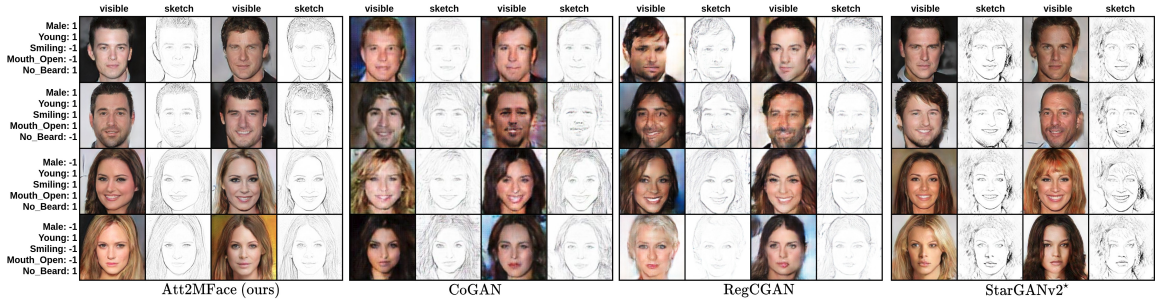


Figure 4.6: Sample 256 × 256 resolution multimodal images generated by different methods using the CelebA-HQ dataset. (Zoom-in for better visualization)

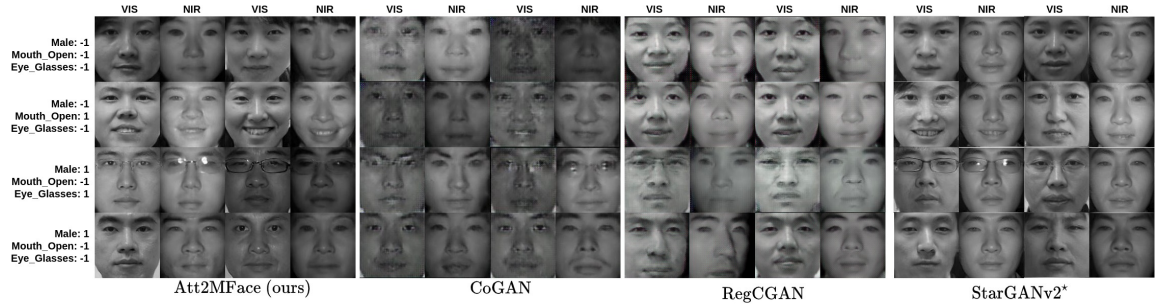


Figure 4.7: Sample 128 × 128 resolution multimodal images generated by different methods using the CASIA NIR-VIS 2.0 dataset. (Zoom-in for better visualization)

are also able to produce reasonable images but the image quality from our model is more photo-realistic than those two methods. Furthermore, the LPIPS distances corresponding to Att2MFace are higher than the other baseline methods which indicates that the generated images from our method are more diverse. The performance of StarGANv2* depends on the quality of the input image which is synthesized by Att2MFace. It also depends on how well StarGANv2* is able to synthesize multimodal images from the visible image. As can be seen from these Figs. 4.5, 4.6, & 4.7 and Table 4.3, in general, StarGANv2* can produce multimodal images but the image quality is poor and the diversity is limited because of the domain discrepancy among different modalities. This experiment clearly shows the significance of our multimodal image generation method.

Additional multimodal synthesized samples corresponding to the proposed method from the ARL Multimodal Face dataset, CelebA-HQ dataset and CASIA-NIR-VIS 2.0 dataset can be found in Figure 4.8, 4.9, and 4.10, respectively.

Attribute Accuracy: In order to check whether the generated images preserve the attributes that were used to synthesize images, we use mean square error (MSE) between the predicted attribute scores corresponding to the reference image and the synthesized image. We fine-tune ResNet50 [103] on the attribute label by replacing the final layer of the classifier. The estimated attributes from the re-trained ResNet50 model are used for evaluation. When conducting this experiment, 5000 ground-truth images with corresponding attributes are randomly selected. The averaged score and standard derivation based on 5 splits are shown in Table 4.5. As can be seen from this table, the proposed Att2MFace is

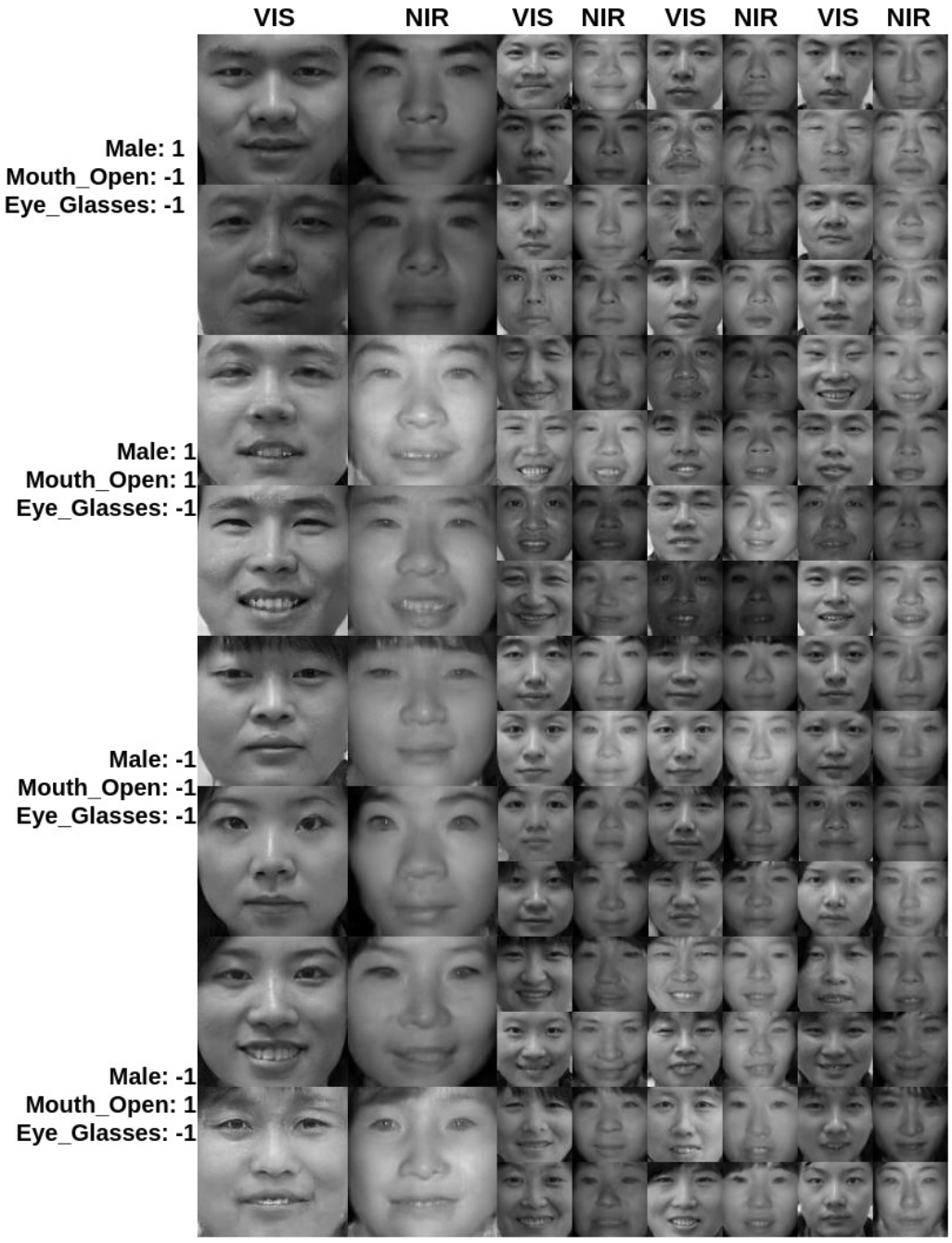


Figure 4.8: Additional 128x128 and 64x64 multimodal images generated using the CASIA-NIR-VIS 2.0 dataset.

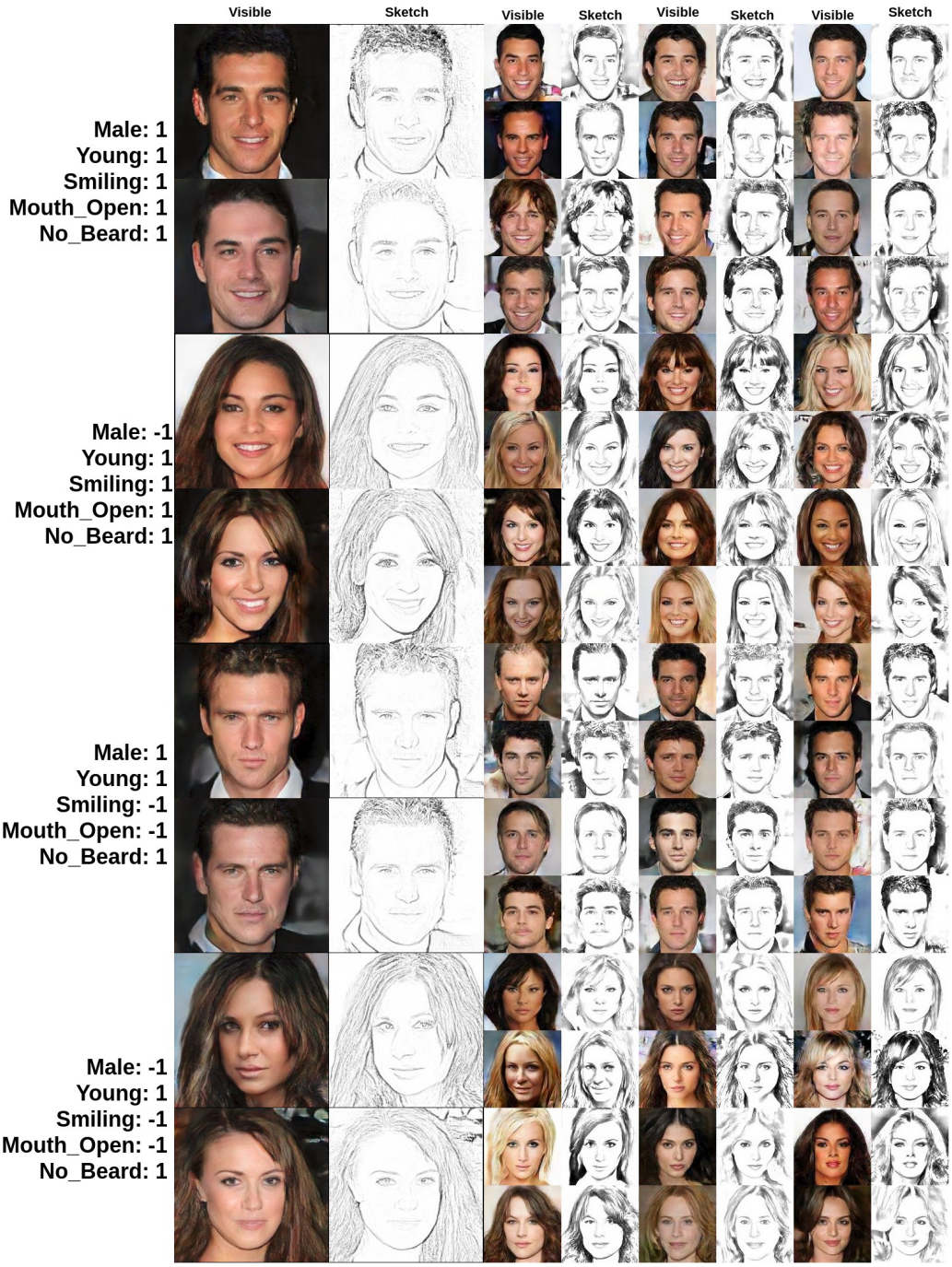


Figure 4.9: Additional 256x256 and 128x128 multimodal images generated using the CELEBA-HQ dataset.

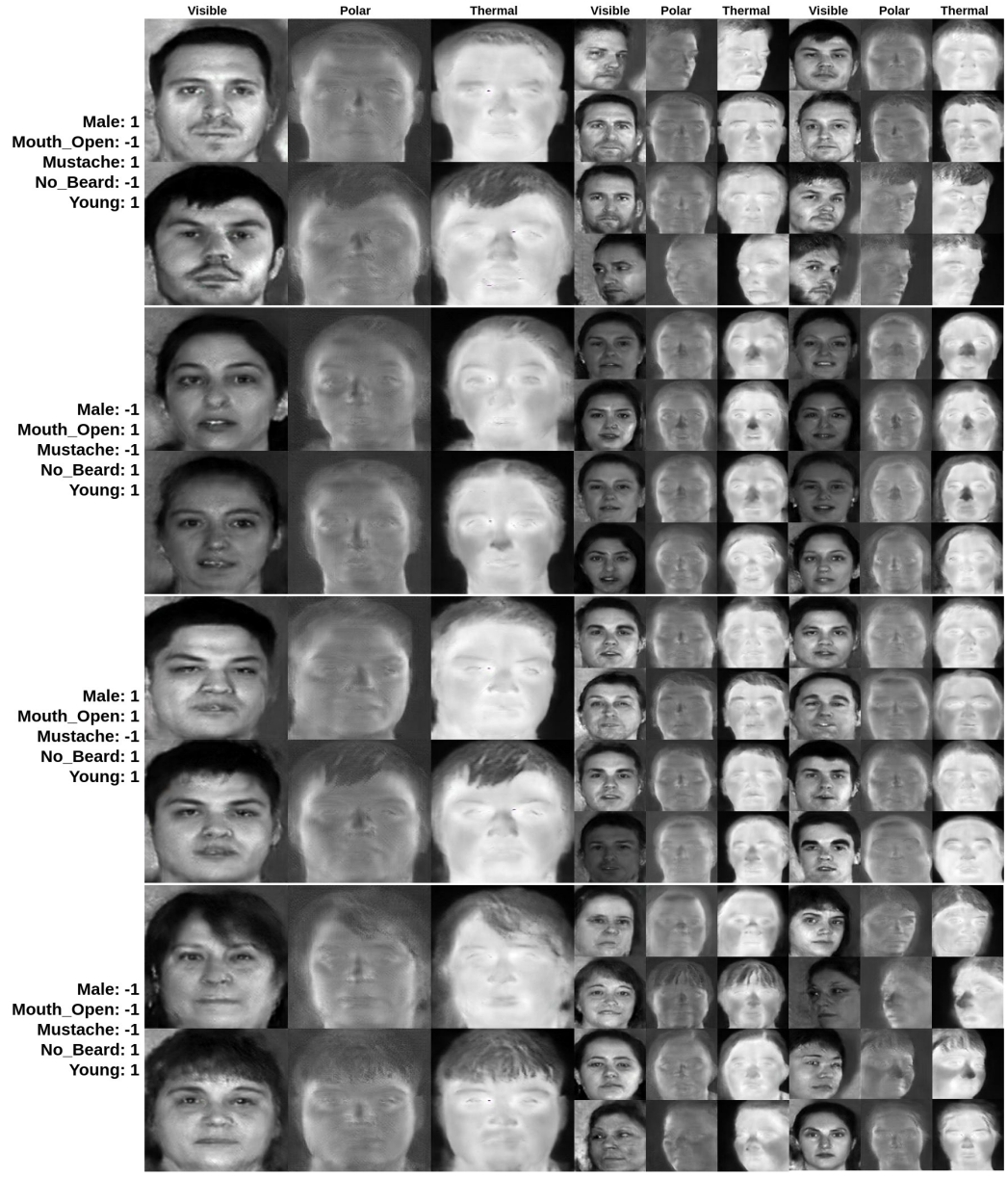
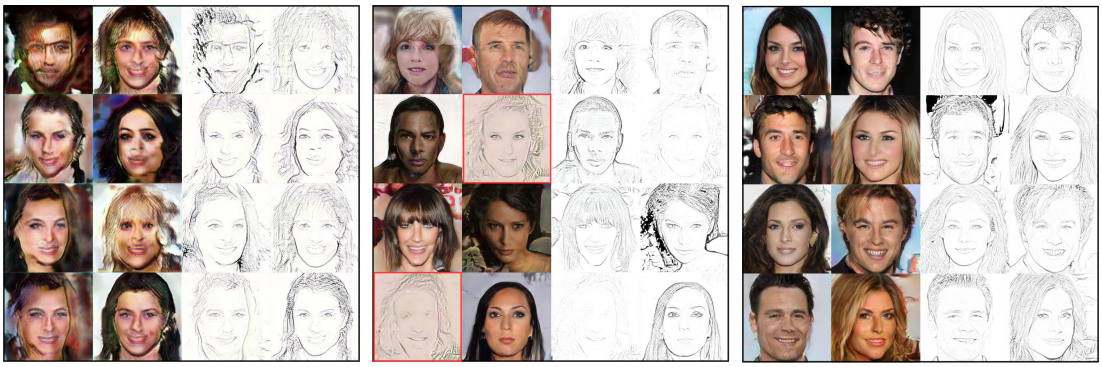


Figure 4.10: Additional 256x256 and 128x128 multimodal images generated using the CASIA-NIR-VIS 2.0 dataset.



(a) Batch Norm (b) Instance Norm (c) Feature Norm

Figure 4.11: Comparison of using different normalization methods. The FID scores for visible/sketch modalities in different normalization methods are: (a) Batch Norm:118.69/96.01; (b) Instance Norm:44.82/28.01; (c) Feature Norm: 13.30/17.75. The modality implications are highlighted with red boxes

able to preserve the attributes on the synthesized images better than the other baselines. Note that since extracting facial attributes from sketch or thermal modalities is difficult, we only estimate attributes from the synthesized visible images.

Normalization Comparison: In order to demonstrate the effectiveness of feature normalization in Eq. (4.2), we compare it with two common normalization methods: batch normalization and instance normalization. For fair comparison, we replace the feature equalization with these two normalizations and keep the other network structure the same. We show the performances visually in Figure 4.11. As can be seen from this figure instance normalization leads to modality implication (highlighted in red boxes). In addition, the FID scores are indicated in the caption to show the superiority of feature equalization quantitatively.

Progressive Learning: Fig. 4.12 shows sample outputs during progressive training of our

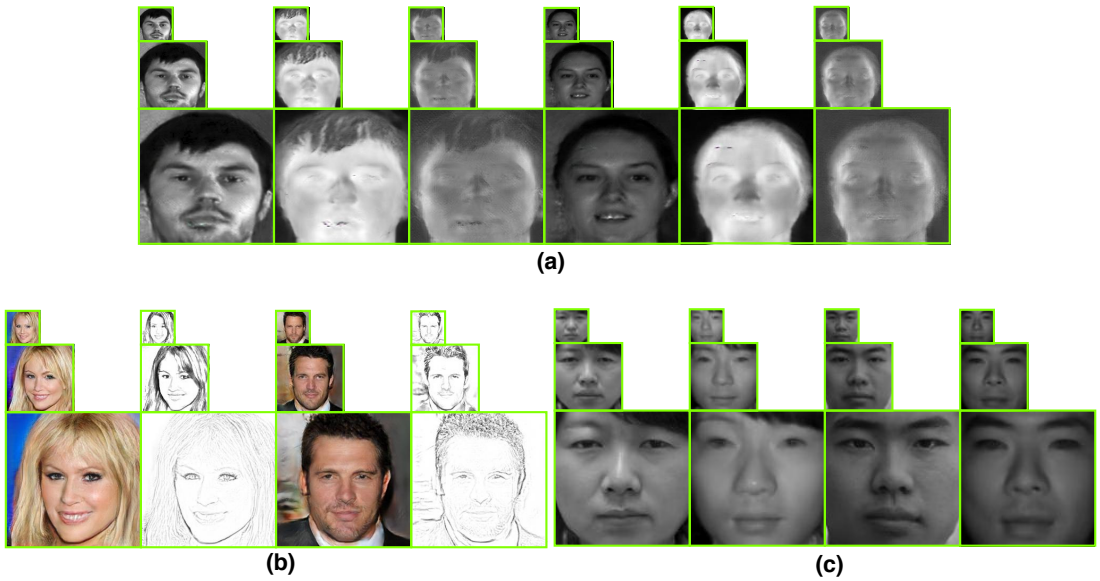


Figure 4.12: Synthesized multimodal images during progressive-growth training at different resolutions.

network. As can be seen from this figure, starting from low-resolution, our method progressively increases the resolution of multimodal images. This incremental learning framework allows the network to discover the overall structure of the face first and then adds finer scale detail at larger scales. Hence, it avoids having to learn multimodal images at all scales simultaneously. This process also stabilizes the training process.

4.2.5 Face Synthesis via Manipulating

To understand whether the model learns to generate a diverse set of images or just memorizes data, two experiments are conducted. In the first experiment, we manipulate the attribute code while keeping the noise vector fixed. This experiment shows the image synthesis capability of our network by manipulating the input attribute. In the second ex-

periment, we fix the attribute code and change the noise vector. This experiment shows whether our model can learn a smooth mapping from noise to the image space.

Attribute Manipulation: Given an attribute code $\mathbf{y}_a$, another attribute code $\mathbf{y}_a^\star$ is obtained by flipping a particular value of $\mathbf{y}_a$ (i.e. Male: -1 $\rightarrow$ Male: 1). The attribute code is interpolated as $\mathbf{y}_a = \beta * \mathbf{y}_a + (1 - \beta) * \mathbf{y}_a$ with $\beta \in [0, 1]$. The manipulated attribute code is then used to synthesize multimodal images by keeping the noise vector fixed. Results corresponding to this experiment are shown in Fig. 4.13. As can be seen from this figure, when higher weights are given to a certain attribute, the corresponding appearance changes. For instance, we are able to synthesize multimodal images with a different gender by changing the weights corresponding to the gender attribute as shown in the first three rows of Fig. 4.13. Each column shows the progression of gender change as the attribute weights are manipulated from -1 to 1. Similarly, the synthesis results when young and mouth open attributes are changed are also shown in Fig. 4.13. It is worth noting that when the attribute weights other than the gender attribute are changed, the identity of the person does not change much.

Noise Manipulation: Similar to the above experiment, we gradually synthesize images from the interpolated latent vectors between the two noise codes while keeping the attribute code frozen. Results corresponding to this experiment are shown in Fig. 4.14. As can be seen from this figure, as we change the noise vector, attributes stay the same while the identity changes. Our model is able to learn the mapping from the latent noise space instead of just memorizing the data.

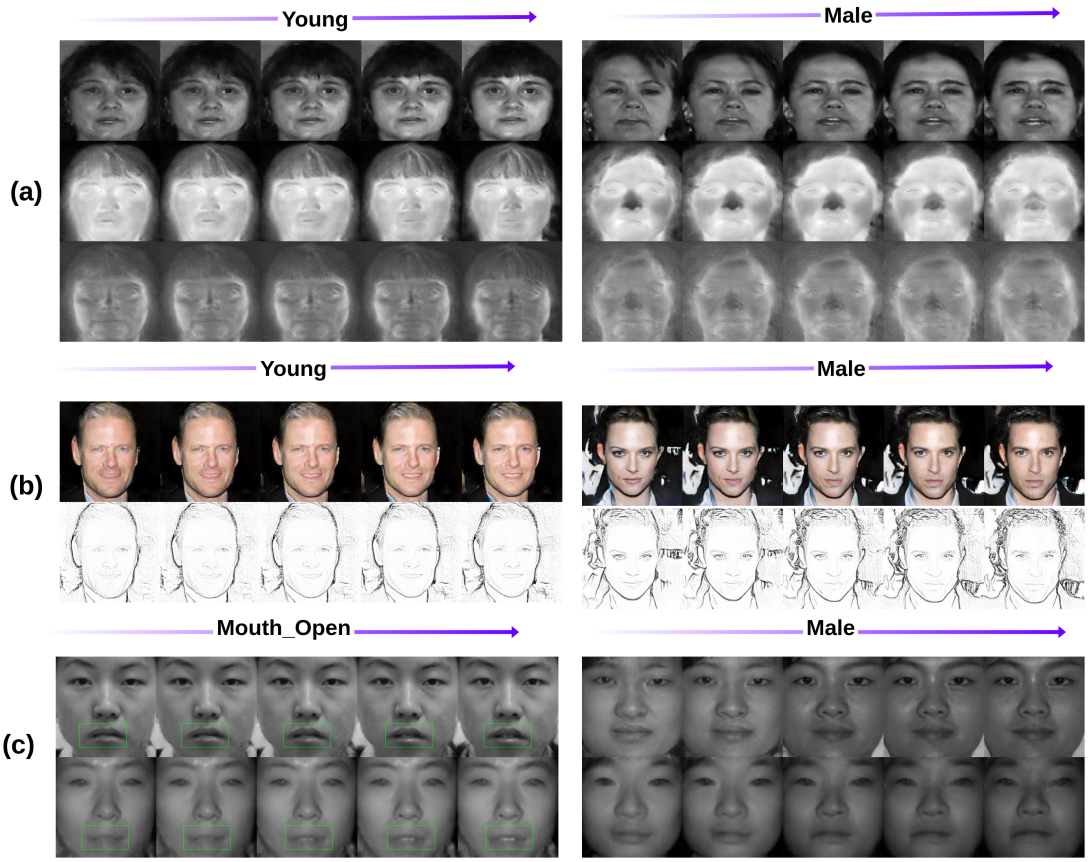


Figure 4.13: Progressive synthesis of multimodal face images when attributes are changed while the noise vector is kept fixed. (a)Old to young and female to male synthesis on the ARL dataset. (b) Old to young and female to male synthesis on the CelebA-HQ dataset. (c) Mouth closed to open and female to male synthesis on the CASIA-NIR-VIS dataset.

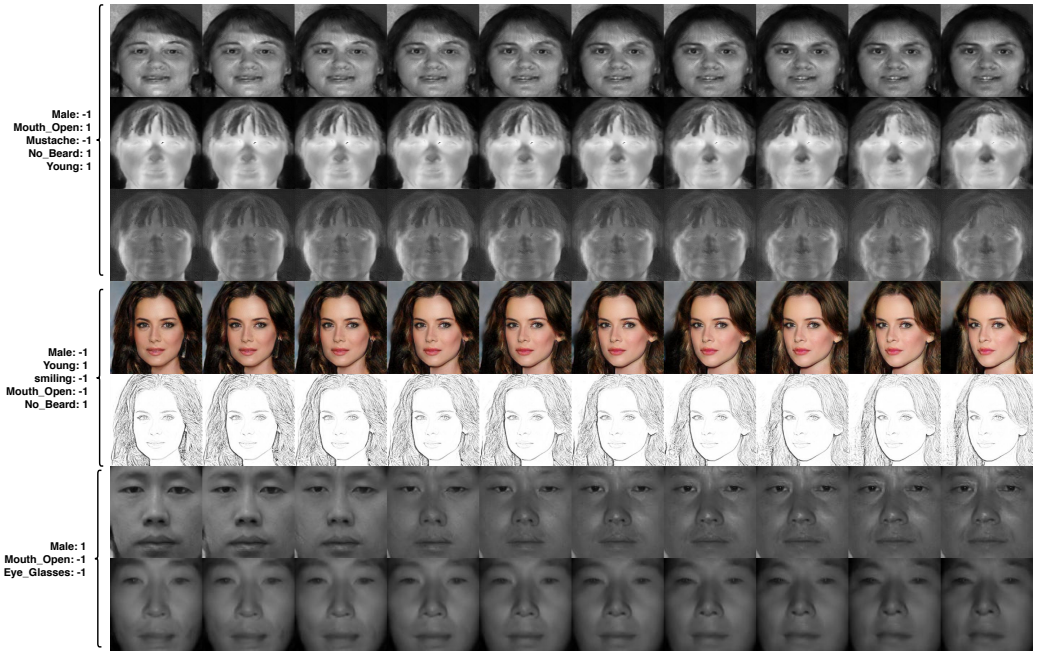


Figure 4.14: Synthesis of multimodal images via interpolation between two noise codes with fixed visual attribute (listed on the left-side). Note that the identity and facial shape change as we vary the noise vector but the attributes are preserved on the synthesized images

4.3 Summary

We proposed a novel network architecture, called Att2MFace, for multimodal face generation from visual attributes. Att2MFace consists of a single generator that can simultaneously generate multimodal face images from visual attributes. Furthermore, we take advance of the progressive training strategy to synthesize consistant multimodal face images. Qualitative and quantitative comparison with other state-of-the-art methods on three datasets demonstrate the superiority of the proposed method. Various experiments showed that the proposed method is able to generate high-quality multimodal images and achieves significant improvements over the state-of-the-art methods.

Table 4.1: The generator network architectures that we use to generate 256x256 multimodal images.

Generator	Act.	Output shape	Block
input code		$(512 + d_a)$	-
Fully-connected	LReLU, Reshape	$512 \times 4 \times 4$	MLP
Conv 3×3	LReLU	$512 \times 4 \times 4$	Initial Block
$c\times$ Conv 1×1		$3 \times 4 \times 4$	Multimodal Stretch-out
Upsample		$512 \times 8 \times 8$	Transition Block
Conv 3×3	LReLU	$512 \times 8 \times 8$	
Conv 3×3	LReLU	$512 \times 8 \times 8$	
$c\times$ Conv 1×1	linear	$3 \times 8 \times 8$	Multimodal Stretch-out
Upsample		$512 \times 16 \times 16$	Transition Block
Conv 3×3	LReLU	$256 \times 16 \times 16$	
Conv 3×3	LReLU	$256 \times 16 \times 16$	
$c\times$ Conv 1×1	linear	$3 \times 16 \times 16$	Multimodal Stretch-out
Upsample		$256 \times 32 \times 32$	Transition Block
Conv 3×3	LReLU	$128 \times 32 \times 32$	
Conv 3×3	LReLU	$128 \times 32 \times 32$	
$c\times$ Conv 1×1	linear	$3 \times 32 \times 32$	Multimodal Stretch-out
Upsample		$128 \times 64 \times 64$	Transition Block
Conv 3×3	LReLU	$64 \times 64 \times 64$	
Conv 3×3	LReLU	$64 \times 64 \times 64$	
$c\times$ Conv 1×1	linear	$3 \times 64 \times 64$	Multimodal Stretch-out
Upsample		$64 \times 128 \times 128$	Transition Block
Conv 3×3	LReLU	$32 \times 128 \times 128$	
Conv 3×3	LReLU	$32 \times 128 \times 128$	
$c\times$ Conv 1×1	linear	$3 \times 128 \times 128$	Multimodal Stretch-out
Upsample		$32 \times 256 \times 256$	Transition Block
Conv 3×3	LReLU	$16 \times 256 \times 256$	
Conv 3×3	LReLU	$16 \times 256 \times 256$	
$c\times$ Conv 1×1	linear	$3 \times 256 \times 256$	Multimodal Stretch-out

Table 4.2: The discriminator network architectures that we use to generate 256x256 multimodal images.

Discriminator	Act.	Output shape	Block
input images		$3 \times 256 \times 256$	-
$c\times$ Conv 1×1	linear	$16 \times 256 \times 256$	Multimodal Stretch-in
Conv 3×3 Conv 3×3 Downsample	LReLU LReLU	$16 \times 256 \times 256$ $32 \times 256 \times 256$ $32 \times 128 \times 128$	Down-stream
$c\times$ Conv 1×1	linear	$32 \times 128 \times 128$	Multimodal Stretch-in
Conv 3×3 Conv 3×3 Downsample	LReLU LReLU	$32 \times 128 \times 128$ $64 \times 128 \times 128$ $64 \times 64 \times 64$	Down-stream
$c\times$ Conv 1×1	linear	$64 \times 64 \times 64$	Multimodal Stretch-in
Conv 3×3 Conv 3×3 Downsample	LReLU LReLU	$64 \times 64 \times 64$ $128 \times 64 \times 64$ $128 \times 32 \times 32$	Down-stream
$c\times$ Conv 1×1	linear	$128 \times 32 \times 32$	Multimodal Stretch-in
Conv 3×3 Conv 3×3 Downsample	LReLU LReLU	$128 \times 32 \times 32$ $256 \times 32 \times 32$ $256 \times 16 \times 16$	Down-stream
$c\times$ Conv 1×1	linear	$256 \times 16 \times 16$	Multimodal Stretch-in
Conv 3×3 Conv 3×3 Downsample	LReLU LReLU	$256 \times 16 \times 16$ $512 \times 16 \times 16$ $512 \times 8 \times 8$	Down-stream
$c\times$ Conv 1×1	linear	$512 \times 8 \times 8$	Multimodal Stretch-in
Conv 3×3 Conv 3×3 Downsample	LReLU LReLU	$512 \times 8 \times 8$ $512 \times 8 \times 8$ $512 \times 4 \times 4$	Down-stream
$c\times$ Conv 1×1		$512 \times 4 \times 4$	Multimodal Stretch-in
Conv 3×3	LReLU Reshape	$512 \times 4 \times 4$ 8192	Down-stream
Fully-connected / Fully-connected		16 / $(d_a + d_c)$	D_A / D_C

Table 4.3: Quantitative results in terms of the FID ↓ (LPIPS ↑) scores corresponding to different methods. Mean value is calculated by averaging the available FID scores from each modality.

	ARL Multimodal Face Database				CelebA Database			CASIA NIR-VIS 2.0		
Methods	Visible	Polarimetric	S0	Mean	Visible	Sketch	Mean	Vis	NIR	Mean
CoGAN [24]	397.22 (0.3942)	275.60 (**0.4647**)	311.55 (**0.4870**)	328.12 (0.4486)	110.92 (0.5053)	113.29 (0.3222)	112.10 (0.4137)	306.03 (0.3103)	97.27 (**0.3494**)	201.65 (0.3298)
RegCGAN [138]	382.75 (0.4792)	264.86 (0.4484)	299.53 (0.4079)	315.71 (0.4451)	108.55 (0.5183)	118.52 (0.3338)	113.53 (0.4260)	142.32 (0.3201)	117.66 (0.3467)	129.99 (0.3334)
StarGANv2* [139]	65.26 (0.4848)	127.17 (0.1078)	157.54 (0.1684)	116.65 (0.2536)	13.30 (0.5494)	115.51 (0.2039)	64.40 (0.3766)	35.03 (0.3815)	54.45 (0.1707)	44.74 (0.2761)
Att2MFace (ours)	**65.26 (0.4848)**	**43.04** (0.4542)	**69.36** (0.4351)	**59.22 (0.4580)**	**13.30 (0.5494)**	**17.75 (0.4568)**	**15.52 (0.5031)**	**35.03 (0.3815)**	**52.64** (0.3343)	**43.83 (0.3579)**

Table 4.4: List of selected visual-attributes.

ARL Multimodal Face Dataset	Male, Mouth_Open, Mustache, No_Bear and Young.
CelebA-HQ	Male, Young, Smiling, Mouth_Open, No_Beard.
CASIA NIR-VIS 2.0	Male, Eye_Glasses, Smiling

Table 4.5: Attribute accuracy based on the MSE metric with mean $\pm$ std.

Methods	ARL Multimodal Face Database	CelebA Database	CASIA NIR-VIS 2.0
CoGAN [24]	0.0583 ± 0.0181	0.0671 ± 0.0152	0.0603 ± 0.0135
RegCGAN [138]	0.0594 ± 0.0146	0.0583 ± 0.0132	0.0574 ± 0.0177
StarGANv2* [139]	0.0547 ± 0.0142	0.0563 ± 0.0094	0.0553 ± 0.0136
Att2MFace (ours)	$\mathbf{0.0516 \pm 0.0158}$	$\mathbf{0.0489 \pm 0.0133}$	$\mathbf{0.0508 \pm 0.0148}$

Chapter 5

Multi-Scale Thermal to Visible Face Verification via Attribute Guided Synthesis

Face Recognition (FR) is one of the most widely studied problems in computer vision and biometrics research communities due to its applications in authentication, surveillance, and security. Various methods have been developed over the last two decades that specifically attempt to address the challenges such as aging, occlusion, disguise, variations in pose, expression, and illumination. In particular, convolutional neural network (CNN) based FR methods have gained significant traction in recent years [144]. This is mainly due to the availability of large annotated datasets, affordability of graphics processing units (GPUs), and trainability of nonlinear layers of deep neural networks employing activation

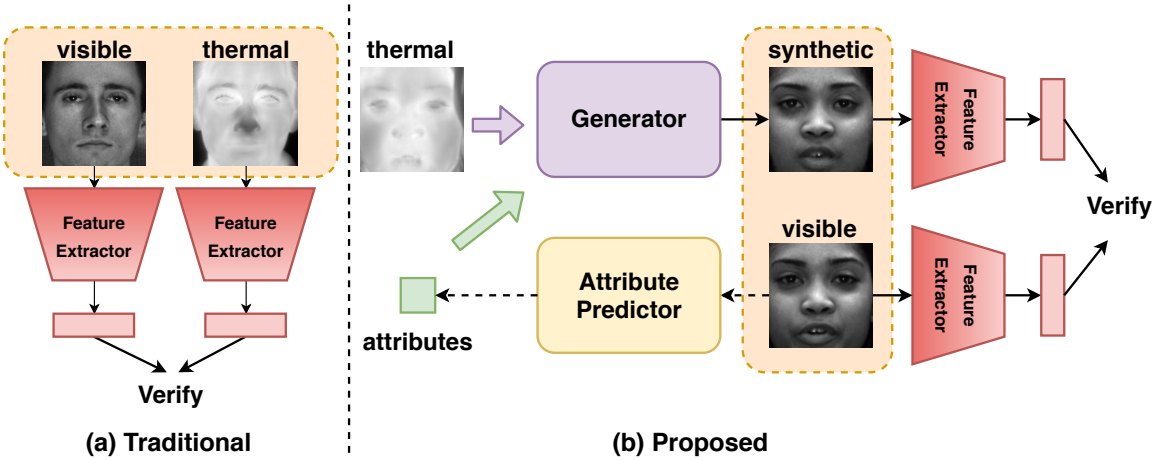


Figure 5.1: (a) Traditional heterogeneous face verification approaches use the features directly extracted from different modalities for verification [1–4]. (b) The proposed heterogeneous face verification approach uses a thermal face and semantic attributes to synthesize a visible face. Then, deep features extracted from the synthesized and visible faces are used for verification.

functions (e.g., ReLU, ELU) that alleviated issues with diminishing/exploding gradients. Many deep CNN-based methods [144–151] have achieved state-of-the-art performances on various FR benchmarks.

Despite the success of CNN-based methods in addressing various challenges in FR, they are fundamentally limited to recognizing face images that are collected near-infrared spectrum. In many practical scenarios such as surveillance in low-light conditions, one has to detect and recognize faces that are captured using thermal modalities [2, 26–28, 140, 152–156]. However, the performance of many deep learning-based methods degrades significantly when they are presented with thermal face images. For example, it was shown in [26,28,63,64] that simply using deep features extracted from both thermal and visible facial images are not sufficient enough for heterogeneous face recognition. The performance

degradation is mainly due to the significant distributional change between the thermal and visible domains as well as a lack of sufficient data for training the deep networks for cross-modal synthesis and matching.

Several attempts have been made to address the thermal-to-visible cross-spectrum FR problem [26–28, 34, 63]. Riggan *et al.* [27] proposed a two-step method (visible feature estimation and visible image reconstruction) to solve the heterogeneous FR problem. Zhang *et al.* [26] proposed a generative adversarial network (GAN) based method that fuses different Stokes images to synthesize a visible face image given the corresponding polarimetric thermal images. Recently, Riggan *et al.* [28] developed a global and local region-based method to improve the discriminative quality of the synthesized visible imagery. Recently, Zhang *et al.* [34] introduced a multi-stream feature-level fusion method to synthesize high-quality visible images from polarimetric thermal images. Though these methods are able to synthesize photo-realistic visible face images to some extent, the synthesized results in [19, 26, 28] are still far from optimal and they tend to lose some semantic attribute information such as expression, facial hair, gender, etc. Such reconstructions may degrade the performance of thermal-to-visible face verification.

In this work, we take a different approach to the problem of thermal-to-visible matching. Fig. 5.1 compares the traditional cross-modal verification problem with that of the proposed attribute-preserved heterogeneous face verification approach. Given a visible and thermal image pair, the traditional approach first extracts some features from these images and then verifies the identity based on the extracted features [2] (see Fig. 5.1(a)). In con-

trast, we propose a novel framework in which we make use of the attributes extracted from the visible image to synthesize the attribute-preserved visible image from the input thermal image for matching (see Fig. 5.1(b)). In particular, a pre-trained VGG-Face model [145] is used to extract the attributes from the visible image. Then, a novel Multi-Scale Attribute Preserved Generative Adversarial Network (Multi-AP-GAN) is proposed to synthesize the visible image from the thermal image guided by the extracted attributes. Finally, a pre-trained VGG-Face network is used to extract features from the synthesized and the input visible images for verification.

5.1 Proposed Method

In this chapter, we discuss the details of the proposed Multi-AP-GAN method (see Fig. 5.2). In particular, we discuss the proposed attribute predictor, multi-scale generator, a series of distinct accompanying discriminators and the loss function used to train these networks.

5.1.1 Attribute Predictor

To efficiently extract attributes from a given visible face, an attribute predictor is fine-tuned based on the VGG-Face network [145] using the annotated attributes. This network is trained separately from Multi-AP-GAN. The fine-tuned network is used in both obtaining the visible face attributes and for capturing the attribute loss when training the generator

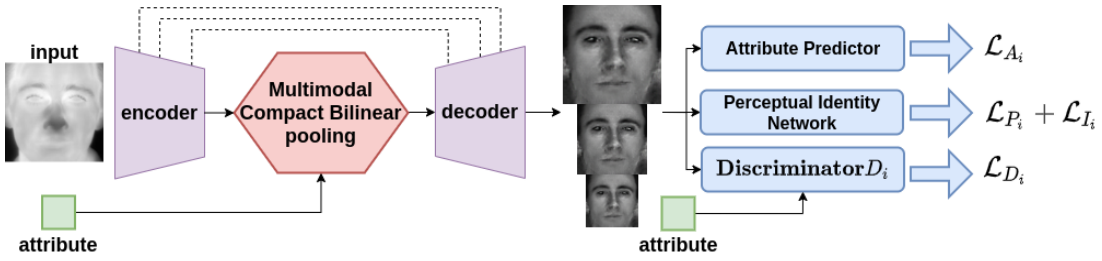
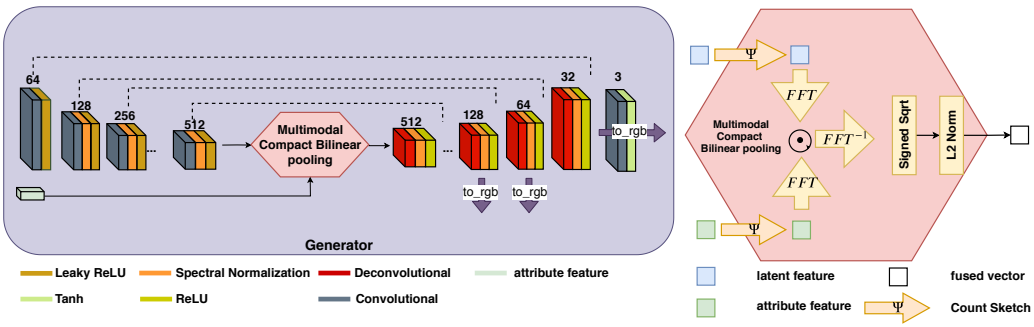


Figure 5.2: A single generator with multi-scale resolution output is proposed to synthesize high-quality images by leveraging hierarchical information at different scales. Multimodal Bilinear Pooling (MCB) pooling is proposed to fuse the semantic attribute information with the image feature in the latent space. In order to make sure that the synthesized image maintains the identity and semantic attributes, a multi-purpose objective function is adopted which consists of adversarial loss $\mathcal{L}_{D_i}$, $\mathcal{L}_1$ loss, perceptual loss $\mathcal{L}_{P_i}$, identity loss $\mathcal{L}_{I_i}$ and attribute preserving loss $\mathcal{L}_{A_i}$.



(a) The multi-scale generator (b) Multimodal Compact Bilinear (MCB) pooling

Figure 5.3: The network architecture of multiscale generator and multimodal compact bilinear (MCB) pooling in details.

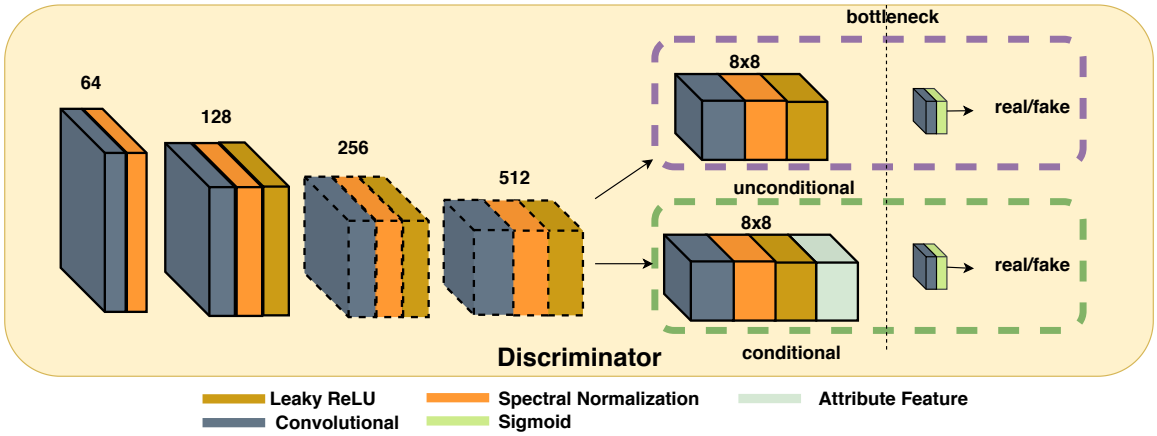


Figure 5.4: An overview of the triplet-pair-input discriminator. The triplet-pair-input discriminator is composed of a conditional and an unconditional streams. The unconditional stream aims to discriminate the fake and real images. The conditional stream aims to discriminate between the image and the corresponding attributes. In order to keep the bottleneck feature map size to be consistent to 8×8 for different input image resolution scale, a different number of downsampling layers (dash-line cubic) are utilized.

and discriminator. When fine-tuning the network, a binary cross-entropy loss is used and the final fully-connected layer has the same dimension as the number of visual attributes. The predictor is selected based on the lowest loss error.

5.1.2 Generator

A U-net structure [29] is used as the building block for the multi-scale generator since it is able to better capture the large receptive field and also able to efficiently address the vanishing gradient problem. In addition, to effectively combine the extra facial attribute information into the building block, we fuse the attribute vector and the image feature in the latent space [19, 26, 107]. Note that the attributes are extracted from the given visible face using the fine-tuned model as discussed above. The generator architecture is illustrated

in Fig 6.3(a).

In our experiments, we observe that simple concatenation of the two vectors (encoded image vector and attribute vector) does not work well. One possible reason is that both vectors are significantly different in terms of their dimensionality. Thus, we adopt the well-known MCB pooling method [32, 33] to overcome this issue. Instead of simple concatenation, MCB leverages the following two techniques: bilinear pooling and sketch count. Bilinear pooling is the outer-product and linearization of two vectors, where all elements of both vectors are interacting with each other in a multiplicative way. In order to overcome the high-dimension computation of bilinear pooling, Pham *et al.* [157] implemented the count sketch of the outer product of two vectors, which involves the Fast Fourier Transform (FFT) and inverse Fast Fourier Transform (FFT^{-1}). The architecture of the MCB module is shown in Fig 6.3(b).

In order to improve the quality of the synthesized visible images, the proposed single generator utilizes a multi-scale output architecture. Specifically, the generator G produces multiple outputs at different resolution scales as follows

$$G(\mathbf{x}, \mathbf{z}) = \{\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_s\}, \tag{5.1}$$

where $\mathbf{x}, \mathbf{z}$ denote the input thermal image and the extracted visual-attributes, respectively. Here, $\{\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_s\}$ denote the synthesized images with gradually growing resolutions and $\hat{\mathbf{y}}_s$ is the final output with the highest resolution s. In this work, we set $s = 3$ where $\hat{y}_3$

Table 5.1: Architecture details corresponding to the generator network.

	conv	conv	conv	conv	conv	conv	conv	MCB	dconv	dconv	dconv	dconv	dconv (to_rgb)	dconv (to_rgb)	dconv (to_rgb)
Input Size	256	128	64	32	16	8	4	2	2	4	8	16	32 (128)	64 (64)	128 (32)
Output Channel	64	128	256	512	512	512	512	512	512	512	512	256	128 (3)	64 (3)	32 (3)
Kernel Size	3	3	3	3	3	3	3	-	3	3	3	3	3 (3)	3 (3)	3 (3)
Stride Size	2	2	2	2	2	2	2	-	2	2	2	2	2 (1)	2 (1)	2 (1)

is the 256×256 image, $\hat{y}_2$ is the 128×128 image, and $\hat{y}_1$ is the 64×64 image. These multi-scale resolution outputs act as a regularizer to the generator G. Furthermore, they shorten the error signal flow path and help to improve the training stability [22].

The multi-scale generator network, as shown in Fig. 6.3(a), consists of the following components:

CL(64)-CNL(128)-CNL(256)-CNL(512)-CNL(512)-CNL(512)-CNL(512)-MCB(512)-DNR(512)-DNR(512)-DNR(512)-DNR(256)-DNR(128)-DNR(64)-DNR(32),

where C stands for the convolutional layer (conv), L stands for LeakyReLU layer (negative_slope=0.02), N stands for the spectral normalization layer [158], MCB indicates the Multimodal Compact Bilinear module [32, 33], D stands for the deconvolutional layer (dconv), and R corresponds to the ReLU layer. All the numbers in parenthesis indicate the channel number of the output feature maps. Table 5.1 gives the details of the generator architecture. Note that, for simplicity, spectral normalization [158], LeakyReLU and ReLU layers are omitted. In the last three layers, feature maps are converted into three-channel images by a "to_rgb" block, which consists of one convolutional layer (parameters are indicated in parenthesis) followed by a Tanh layer.

5.1.3 Discriminator

A series of distinct discriminators $D_i, i = 1, \cdots, s$ are utilized and trained iteratively with the generator G. For a certain discriminator at the i-th resolution scale, a patch-based discriminator [159] is leveraged and it not only aims to discriminate between real/fake images but also to discriminate between the image and the corresponding attributes. Similar to previous works [19,21,22], a triplet of paired image and attribute is given to the discriminator: $real$, $fake$ and $wrong$. The $real$ pair consists of a real-image ($\mathbf{y}_i$) along with the corresponding true-attributes ($\mathbf{z}$). The $wrong$ pair consists of a real image ($\mathbf{y}_i$) along with wrong attributes ($\mathbf{z}'$). The $fake$ pair consists of a fake-image ($\hat{\mathbf{y}}_i$) with true attributes ($\mathbf{z}$). The overall adversarial objective function used to train the network is as follows:

$$\begin{aligned}
\mathcal{L}_G &= \sum_{i=1}^{s} \min_G \max_{D_i} (V^i_{real} + V^i_{fake} + V^i_{wrong}), \\
V^i_{real} &= \mathbb{E}_{\mathbf{y}_i \sim P_Y}[\log D_i(\mathbf{y}_i)] \\
&\quad + \mathbb{E}_{\mathbf{y}_i, \mathbf{z} \sim P_{Y,Z}}[\log D_i(\mathbf{y}_i, \mathbf{z})] \\
V^i_{wrong} &= \mathbb{E}_{\mathbf{y}_i, \mathbf{z}' \sim P_{Y,Z}}[\log(1 - D_i(\mathbf{y}_i, \mathbf{z}'))] \\
V^i_{fake} &= \mathbb{E}_{\hat{\mathbf{y}}_i \sim P_{G(\mathbf{x},\mathbf{z})}}[\log(1 - D_i(\hat{\mathbf{y}}_i))] \\
&\quad + \mathbb{E}_{\hat{\mathbf{y}}_i \sim P_{G(\mathbf{x},\mathbf{z})}, \mathbf{z} \sim P_Z}[\log(1 - D_i(\hat{\mathbf{y}}_i, \mathbf{z}))].
\end{aligned} \tag{5.2}$$

Specifically, each discriminator D_i has two streams: conditional stream and unconditional stream. One discriminator on 256×256 resolution scale is illustrated in Fig. 6.4. The unconditional stream aims to learn the discrimination between the real and the synthesized

images. This unconditional adversarial loss is back-propagated to G to make sure that the generated samples are as realistic as possible. In addition, the conditional stream aims to learn whether the given image matches the given attributes or not. This conditional adversarial loss is back-propagated to G so that it generates samples that are attribute-preserving.

Fig. 6.4 gives an overview of a discriminator at 256×256 resolution scale. This discriminator consists of 6 convolutional blocks for both conditional and unconditional streams. Details of these convolutional blocks are as follows:

CL(64)-CNL(128)-CNL(256)-CNL(512)-C^{2}NL(512)-C^{2}S(1),

where S stands for the Sigmoid activation layer. Note that the only difference between the unconditional and conditional stream is the concatenation of the attribute vector at the fifth convolutional block. For different discriminator, D_i at different resolution scale, the number of convolutional down-sample blocks (blocks with dotted lines in Fig. 6.4) vary, but we keep the bottleneck feature map at the same size (i.e. 8×8). The architecture details corresponding to the other discriminators are given in Table 5.2.

5.1.4 Loss Function

The generator is optimized by minimizing the following loss

$$\mathcal{L}_{Multi-AP-GAN} = \mathcal{L}_G + \lambda_A \mathcal{L}_A + \lambda_P \mathcal{L}_P + \lambda_I \mathcal{L}_I + \lambda_1 \mathcal{L}_1, \tag{5.3}$$

[2]unconditional and conditional streams are shortened for brevity.

Table 5.2: Architecture details corresponding to different discriminators. Numbers in parenthesis indicate the channel number of the output feature maps. The convolutional layers have stride size 2.

Discriminator 64x64	Discriminator 128x128	Discriminator 256x256
Convolutional (64) LeakyReLU	Convolutional (64) LeakyReLU	Convolutional (64) LeakyReLU
Convolutional (128) Spectral Norm LeakyReLU	Convolutional (128) Spectral Norm LeakyReLU	Convolutional (128) Spectral Norm LeakyReLU
Convolutional^2 (256) Spectral Norm LeakyReLU	Convolutional (256) Spectral Norm LeakyReLU	Convolutional (256) Spectral Norm LeakyReLU
Convolutional^2 (1) Sigmoid	Convolutional^2 (512) Spectral Norm LeakyReLU	Convolutional (512) Spectral Norm LeakyReLU
	Convolutional^2 (1) Sigmoid	Convolutional^2 (512) Spectral Norm LeakyReLU
		Convolutional^2 (1) Sigmoid

where $\mathcal{L}_G$ is the multi-scale adversarial loss in Eq (5.2) , $\mathcal{L}_P$ is the perceptual loss, $\mathcal{L}_I$ is the identity loss, $\mathcal{L}_A$ is the attribute loss, $\mathcal{L}_1$ is the loss based on the L_1-norm between the target and the reconstructed image, and $\lambda_P, \lambda_I, \lambda_A, \lambda_1$ are the corresponding weights.

5.1.4.1 Multi-scale Perceptual and Identity Loss

Perceptual loss was originally introduced by Johnson *et al.* [160] for style transfer and super-resolution. It has been observed that the perceptual loss produces visually pleasing results than L_1 or L_2 loss. The perceptual and identity losses are defined as follows

$$\mathcal{L}_{P,I} = \sum_{i=1}^{s} \sum_{c=1}^{3} \sum_{w=1}^{W} \sum_{h=1}^{H} \| F(\hat{\mathbf{y}}_i)^{c,w,h} - F(\mathbf{y}_i)^{c,w,h} \|_1, \tag{5.4}$$

where F represents a non-linear CNN feature. VGG-16 [161] is used to extract features in this work. C, W, H are the dimensions of features from a certain level of the VGG-16, which are different for perceptual and identity losses. Since the deeper convolutional layer captures more semantic information, we choose deeper convolutional feature maps as the identity loss.

In addition, multi-scale L_1 loss between the synthesized image $\hat{\mathbf{y}}_i$ and the corresponding real image $\mathbf{y}_i$ is used to capture the low-frequency information, which is defined as follows

$$\mathcal{L}_1 = \sum_{i=1}^{s} \| \hat{\mathbf{y}}_i - \mathbf{y}_i \|_1. \tag{5.5}$$

5.1.4.2 Multi-scale Attribute Loss

Inspired by the perceptual loss, we define an attribute preserving loss, which measures the error between the attributes of the synthesized image and the real image. To make sure the pre-trained model captures the facial attribute information, we fine-tune the pre-trained VGG-Face network on the annotated attribute dataset and regard the fine-tuned attribute classifier as the pre-trained model for the attribute preserving loss. Similar to the perceptual loss, the $\mathcal{L}_A$ is defined as follows

$$\mathcal{L}_A = \sum_{i=1}^{s} \|Q(\hat{\mathbf{y}}_i) - Q(\mathbf{y}_i)\|_1, \tag{5.6}$$

where Q is the fine-tuned attribute predictor network. The output vectors are from the last layer. As a result the feature dimensions C, W, H are omitted in (5.6). By feeding such attribute information into the generator during training, the generator G is able to learn semantic information corresponding to the face.

5.1.5 Implementation

The entire network is trained in Pytorch on a single Nvidia Titan-X GPU. During the Multi-AP-GAN training, the L_1, perceptual and identity loss parameters are chosen as $\lambda_1 = 10$, $\lambda_P = 2.5$, $\lambda_I = 0.5$, respectively. The ADAM [110] is implemented as the optimization algorithm with parameter $betas = (0.5, 0.999)$ and batch size is set equal to 1. The total epochs are 200. For the first 100 epochs, we fix the learning rate as 0.0002 and

for the remaining 100 epochs, the learning rate was decreased by $1/100$ after each epoch. The feature maps for the perceptual and the identity loss are from the relu1-1 and the relu2-2 layers, respectively. In order to fine-tune the attribute predictor network, we manually annotate images with the attributes tabulated in Table 5.3.

Table 5.3: The facial attributes used in this work.

attributes	Arched_Eyebrows, Big_Lips, Big_Nose, Bushy_Eyebrows, Male, Mustache, Narrow_Eyes, No_Beard, Mouth_Slightly_Open, Young

5.2 Datasets and Protocols

In this section, we describe the datasets and the protocols that we use to conduct experiments. In particular, we describe the new extended ARL Polarimetric thermal face dataset and the corresponding protocol that we use in this work.

5.2.1 Extended Polarimetric Thermal Face Dataset

In many recent approaches, the polarization-state information of thermal emissions has been used to achieve improved cross-spectrum face recognition performance [26–28, 140, 152] since it captures geometric and textural details of faces that are not present in the conventional thermal facial images [140, 152]. A polarimetric thermal image consists of three Strokes images: S_0, S_1, S_2 where S_0 indicates the conventional total intensity thermal image, S_1 captures the horizontal and vertical polarization-state information, S_2 captures

the diagonal polarization-state information [140]. Similar to [26,28], we also refer to Polar as the three channel polarimetric image concatenated with S_0 , S_1 and S_2. These Stokes images along with the visible and the polarimetric images corresponding to a subject in the ARL dataset [140] are shown in Fig. 5.5. It can be observed that S_1, S_2 tend to preserve more textural details compared to S_0.

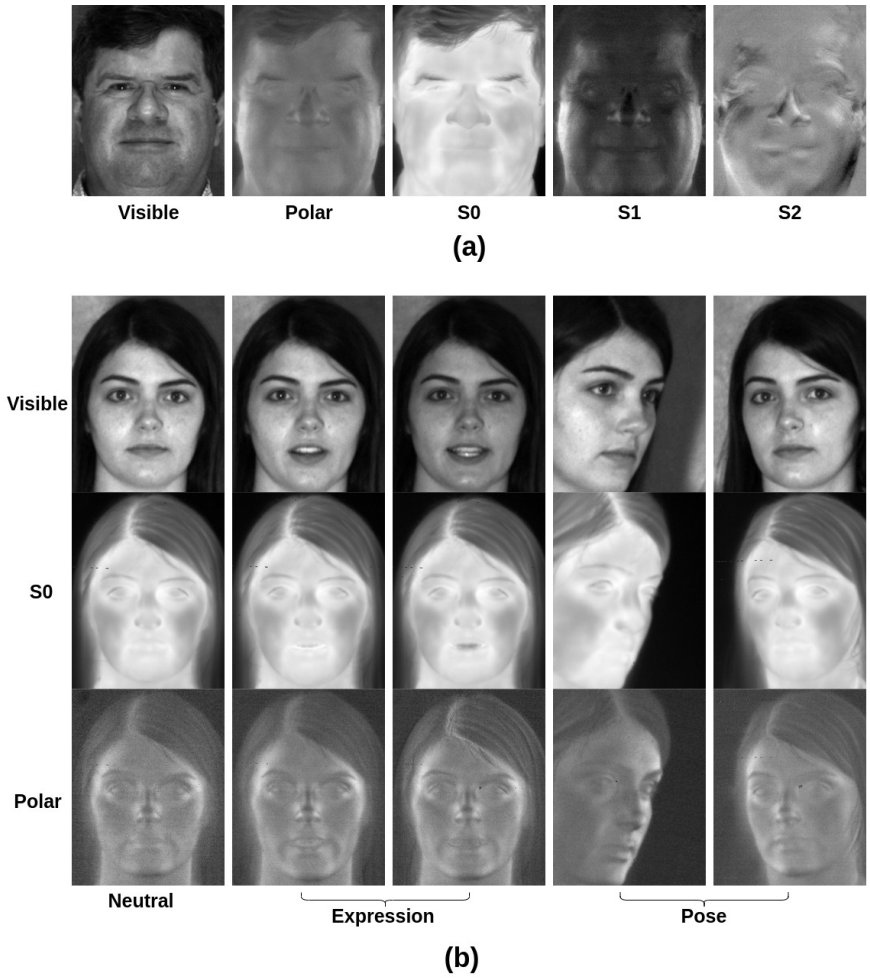


Figure 5.5: Sample images from the ARL dataset. (a) Visible, polarimetric thermal, and Stokes images (S0, S1, S2) corresponding to a subject from the ARL dataset [140]. (b) Sample visible, conventional thermal and polarimetric thermal images with different variations from the ARL Dataset Volume III.

CHAPTER 5. MULTI-SCALE THERMAL TO VISIBLE FACE VERIFICATION VIA ATTRIBUTE GUIDED SYNTHESIS

The U.S. Army DEVCOM Army Research Laboratory (ARL) multimodal face dataset consists of polarimetric thermal and visible face image pairs in three volumes. Volume I consists of the polarimetric thermal and visible images from 60 subjects, which were collected by the U.S. Army Research Laboratory in 2014-2015. Frontal imagery with different ranges and expressions are included. Details regarding this volume can be found in [140] and [34]. Volume II consists of images from 51 subjects collected at a Department of Homeland Security test facility. As described in [34], while the participants of the Volume I subset consisted exclusively of the ARL employees, the participants of the Volume II collect were recruited from the local community in Maryland, resulting in more demographic diversity. In addition, frontal imagery with various expressions is included in this volume.

In this work, we present an extension of the dataset which was collected by ARL across 11 different sessions over 6 days. We refer to this extended dataset as Volume III hereinafter. Volume III contains polarimetric thermal and visible facial signatures from 121 subjects collected at Johns Hopkins University Applied Physics Laboratory as part of an IARPA government testing event. There are a total of 5419 polarimetric thermal and visible image pairs with significant variations (Fig. 5.5) such as expression, off-pose, glasses, etc. These variations make the dataset more challenging for cross-modal face verification. Note that this extended database is available upon request.

To be consistent with previous methods [34, 140], the experimental protocols are defined as follows:

Protocol I: The Protocol I is evaluated on Volume I, which consists of frontal imagery with

range and expression variations (including neutral expression). Images from 30 subjects with eight samples for each subject are used as the training split. Images from the other 30 subjects with eight samples for each subject are used as the test split. All the samples in training and test split are randomly chosen from 60 subjects. Results are evaluated on five random splits. Note that there are no overlapping subjects between training and test splits.

Protocol II: The Protocol II is evaluated on the extended 111 subject dataset which contains the images from both Volume I and Volume II. In particular, 85-subject images are used as the training split and the other 26-subject images are denoted as the test split. The 85-subject images in training split consist of all 60-subject images in Volume I and another 25-subject images randomly selected from Volume II. The other 26-subject images in Volume II are selected as the test split. As before, results are evaluated on five random splits [34]. Note that Volume II consists of frontal imagery with expression variations only (including neutral expression).

Protocol III: The Protocol III is evaluated only on the Volume III data consisting of images from 121 subjects. Volume III includes frontal and off-pose imagery (excludes extreme pose, e.g. profile), and expression variation (including neutral expression). Images from 96 randomly chosen subjects are used as the training split and the images from the remaining 25 subjects are used as the test split. Results are evaluated on five random splits.

5.2.2 Visible and Thermal Paired Face Database

In addition to the ARL dataset, the proposed method is evaluated on a recently introduced Visible and Thermal Paired Face Database [162]. This dataset contains thermal and visible image pairs corresponding to 50 subjects. Each subject participated in two different sessions separated by a time interval of 3 to 4 months. This dataset includes 21 face images per subject in each session. These images correspond to different facial variations in illumination, head pose, expression and occlusion. In total, 4200 images are included in this dataset.

Protocol: Images corresponding to randomly chosen 30 subjects are used as the training split and the images from the remaining 20 subjects are used as the test split. This results in 630 paired training images and 420 paired testing images. There is no overlap among subjects in the training and the test sets. Results are evaluated on five random splits.

5.2.3 Tufts Face Database

We also evaluate the proposed method on a recently proposed Tufts Face Database [37], which contains 1532 paired visible and thermal face images from 112 subjects. For each subject multiple images are taken in different conditions. In particular, each subject has images in 9 different poses, 4 expressions and 1 occlusion with eye glasses. Sample images from this dataset are shown in 5.6. The Tufts dataset [37] is more difficult than the other two datasets as it contains less number of images per person in each variation.

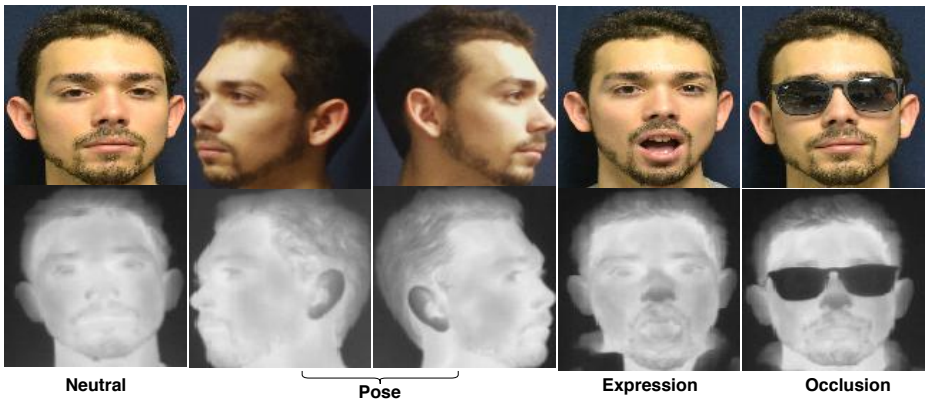
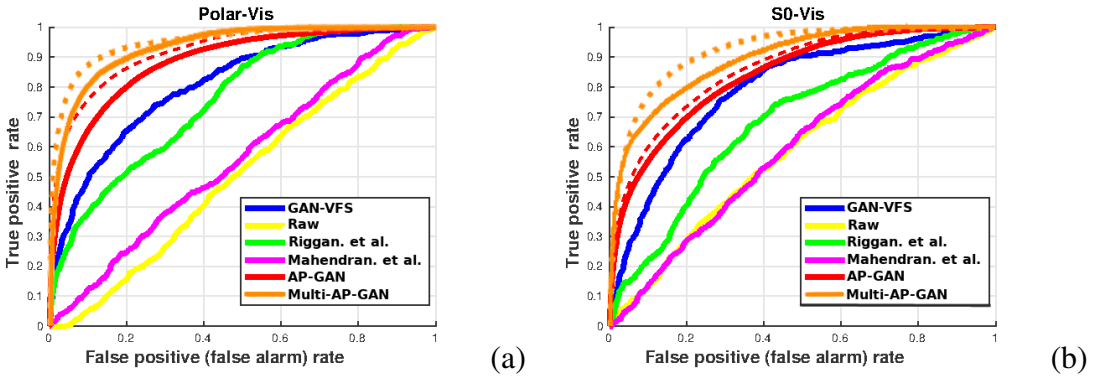


Figure 5.6: Sample thermal and visible images from the Tufts Face Database [37] with different variations.

Protocol: Similar to the previous protocols, images corresponding to 90 subjects are used for training and the images from the remaining 22 subjects are used for testing. This results in about 1232 paired data for training and 300 paired data for testing. There is no overlap among subjects in the training and the test sets. Results are reported based the evaluations on five random splits.

5.2.4 Preprocessing

In addition to the standard preprocessing, two more preprocessing steps are used for the proposed method. First, the faces in the visible images are detected by MTCNN [163]. Then, a standard central crop method is used to crop the detected faces. Since MTCNN is implementable on the visible images only, we use the same detected rectangle coordinations to crop the thermal images, which were already aligned to the same canonical coordinates as the visible images. After preprocessing, all the images are scaled and saved



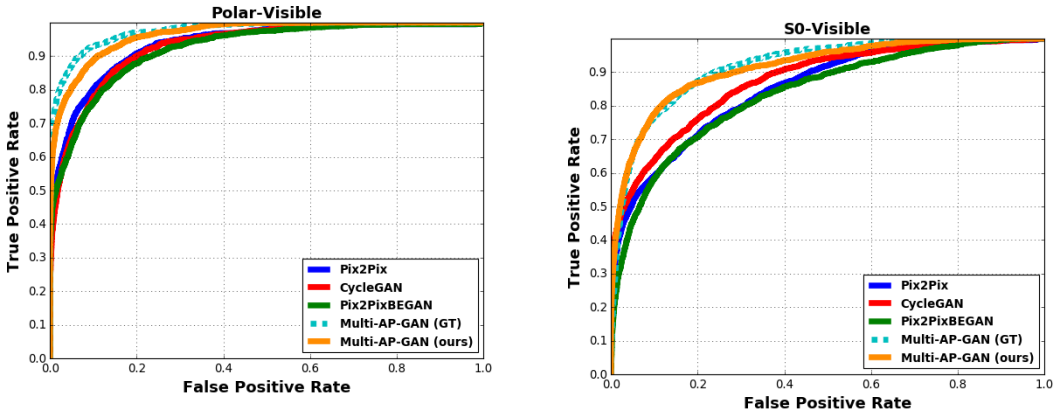
Polarimetric thermal-to-visible verification Thermal-to-visible verification

Figure 5.7: The ROC curve comparison on Protocol I with several state-of-the-art methods: (a) Polarimetric thermal-to-visible verification performance. (b) S0-to-Visible verification performance. Note that the dotted lines indicate results based on the ground-truth attributes. The gap between the results with ground-truth attributes and that with predicted attributes demonstrate the degradation caused by the attribute predictor.

as 256×256 16-bit PNG files.

5.2.5 Metrics

Once the visible image is synthesized from the input probe thermal image, we use a pre-trained VGG-Face model [145] to extract features from the synthesized visible probe image as well as the visible gallery image to perform cross-modal face verification. In particular, the verification score is calculated using the cosine similarity between the two feature vectors. The cross-modal verification performance of different methods is evaluated using the Receiver Operating Characteristic (ROC) curve, Area Under the Curve (AUC) and Equal Error Rate (EER) measures.



(a) Polarimetric thermal-to-visible verification

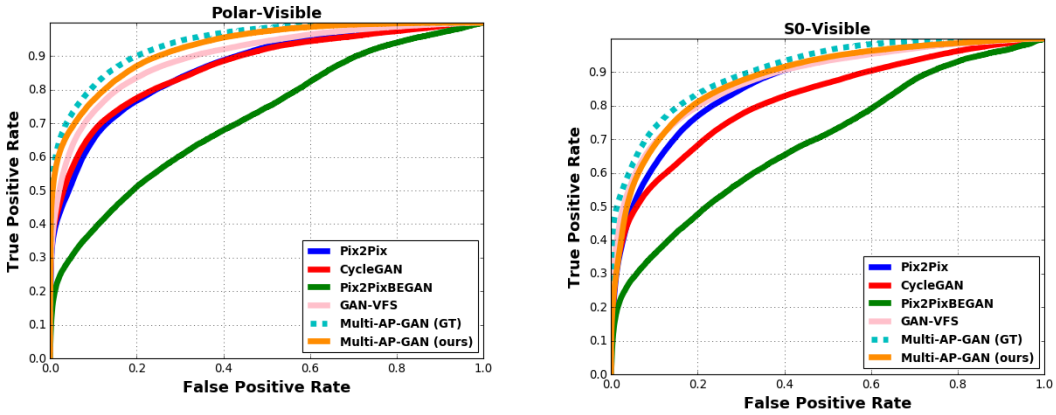
(b) Thermal-to-visible verification

Figure 5.8: The ROC curve comparison on Protocol II with several state-of-the-art methods: (a) Polarimetric thermal-to-visible verification performance. (b) S0-to-Visible verification performance. Note that the dotted lines indicate results based on the ground-truth attributes. The gap between the results with ground-truth attributes and that with predicted attributes demonstrate the degradation caused by the attribute predictor.

5.3 Experimental Results

In this part, we demonstrate the effectiveness of the proposed approach by conducting various experiments on the datasets described in the previous parts. Since the ARL Dataset contains both conventional thermal (S_0) and polarimetric thermal modalities, we conduct the following two cross-modal face verification experiments on the ARL dataset: 1) Conventional thermal (S0) to Visible (Vis) and 2) Polarimetric thermal (Polar) to Visible (Vis). On the other hand, the Visible and Thermal Paired Face Database and the Tufts Face Database do not contain polarimetric thermal images. As a result, we only conduct thermal-to-visible cross-domain face verification experiments on these datasets.

We evaluate and compare the performance of the proposed method with that of the following recent state-of-the-art methods [26–28, 34, 63, 68, 69, 164]. Note that our previous



(a) Polarimetric thermal-to-visible verification

(b) Thermal-to-visible verification

Figure 5.9: The ROC curve comparison on Protocol III with several state-of-the-art methods: (a) Polarimetric thermal-to-visible verification performance. (b) S0-to-Visible verification performance. Note that the dotted lines indicate results based on the ground-truth attributes. The gap between the results with ground-truth attributes and that with predicted attributes demonstrate the degradation caused by the attribute predictor.

work [63] can be viewed as a single scale version of the proposed method. In particular, in [63], we synthesize images at a particular scale which has the same resolution as the input. We also conduct experiments with another baseline method called, Multi-AP-GAN (GT), where we use the ground-truth attributes in our method rather than automatically predicting them using the proposed attribute predictor. This baseline will clearly determine how effective the proposed attribute predictor is in determining the attributes from unconstrained visible faces.

5.3.1 Results on the ARL Face Dataset

Fig. 5.7 shows the performance corresponding to Protocol I on two different experimental settings (i.e S0-to-visible and Polar-to-visible). Compared with other state-of-the-

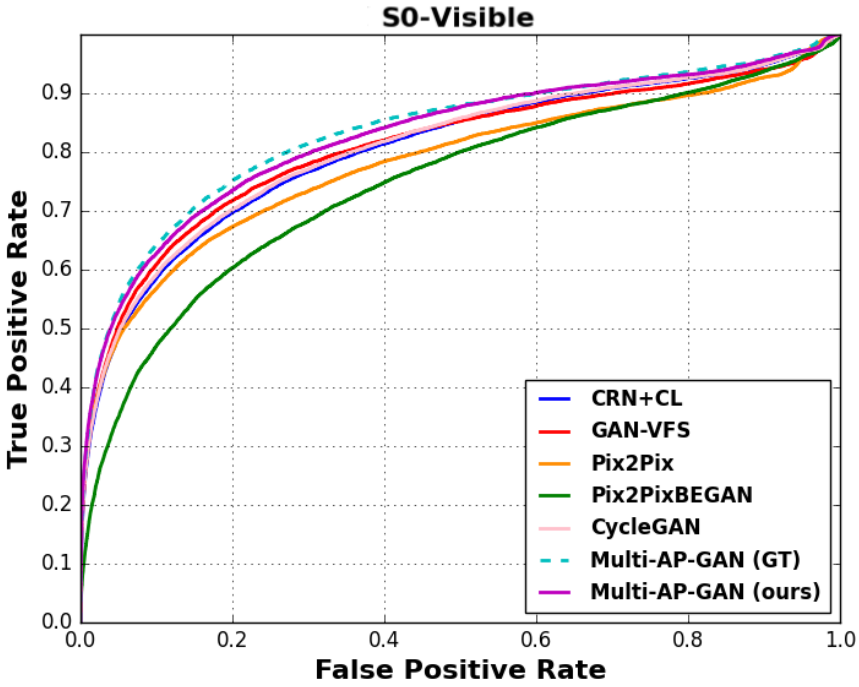


Figure 5.10: The ROC curve comparison on Thermal-Visible Paired Database [162]. Note that the dotted lines indicate results based on the ground-truth attributes. Similarly, the gap between the results with ground-truth attributes and that with predicted attributes demonstrate the degradation caused by the attribute predictor.

art methods in Fig. 5.7, the proposed method performs better with a larger AUC and lower EER scores. In addition, it can be observed that the performance corresponding to the Polar modality is better than the S0 modality, which also demonstrates the advantage of using the polarimetric thermal images than the conventional thermal images. In addition, the gap between the results with ground-truth attributes (dash-line) and that with the predicted attributes (solid-line) demonstrates the degradation caused by the attribute predictor. The quantitative comparisons, as shown in the Table 7.1, also demonstrate the effectiveness of the proposed method. In addition, compared with the previous single scale resolution

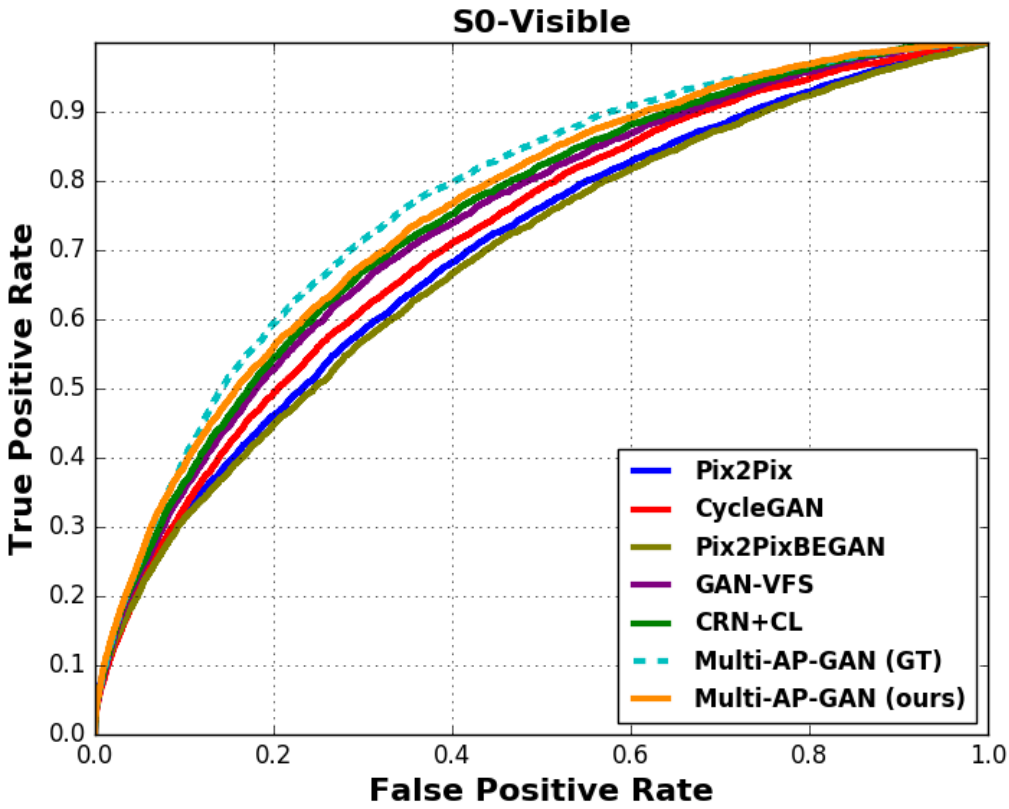


Figure 5.11: The ROC curve comparison on the Tufts Face Database [37]. Note that the dotted lines indicate results based on the ground-truth attributes. Similarly, the gap between the results with the ground-truth attributes and that with predicted attributes demonstrate the degradation caused by the attribute predictor.

method [63], the proposed multi-scale algorithm achieves significant improvement: around 4% and 6% on the conventional and polarimetric thermal modalities, respectively. These improvements demonstrate the effectiveness of the proposed multi-scale synthesis algorithm.

Furthermore, we also show some visual comparisons in Fig. 5.12. The first row in Fig. 5.12 shows one synthesized sample using S0. The second row shows the same synthesized sample using a polarimetric thermal image. It can be observed that the results

Input Riggan *et al.* [27] Mahendran *et al.* [164] GAN-VFS [26] AP-GAN [63] Multi-AP-GAN Ground Truth

Figure 5.12: The visual comparison of synthesized samples from different methods: Riggan *et al.* [27], Mahendran *et al.* [164], GAN-VFS [26], AP-GAN [63], Multi-AP-GAN, Ground Truth. The first row results correspond to the S0 image, and the second row results correspond to the Polar image.

of Riggan *et al.* [27] do capture the overall face structure but it tends to lose some facial details. Results of Mahendran *et al.* [164] are poor compared to [27]. Results of Zhang *et al.* [26] are more photo-realistic but tend to lose some attribute information. The proposed Multi-AP-GAN not only generates photo-realistic images but also preserves attributes on the reconstructed images.

Fig. 5.8 and Table 6.2 show the performance of different methods on Protocol II. These results also demonstrate the superiority of the proposed method. Note that the performance of many methods is slightly better in Protocol II than Protocol I. This is mainly due to the fact that the training dataset is larger in Protocol II than Protocol I.

Protocol III results corresponding to different methods are shown in Fig. 5.9 and Table 5.6. Note that face images in this volume include many variations such as expression, pose, illuminations and occlusion (glasses). As a result, the performance of the methods compared is slightly lower than what we observed in Protocol I and Protocol II. In general,

Input Pix2Pix [159] CycleGAN [165] GAN-VFS [26] CRN+CL [68,69] Multi-AP-GAN Ground Truth

Figure 5.13: The visual comparison of synthesized images corresponding to Pix2Pix [159], CycleGAN [165], GAN-VFS [26], CRN+CL [68,69], Multi-AP-GAN (ours) from the Visible and Thermal Paired Face Database [162].

Pose-Up Pose-Left Pose-Right Pose-Down Occolusion-Hand on Eye Light-Dark

Figure 5.14: Some failure cases. Note that extreme pose, illumination and occlusion variations cause the proposed method to synthesize poor quality images.

the proposed method performs favorably against the state-of-the-art methods. Note that Pix2PixBEGAN method [159,166] fails to generate good quality visible faces from profile thermal face images. As a result, Pix2PixBEGAN method performs poorly on this dataset.

We further analyze the cross-modal verification performance of different methods on different variation settings on Protocol III. The corresponding results are shown in Table 5.7. Since variations like occlusion and illumination are not included in some subjects, we only use three variations (neutral, expression, and pose) which are included in all subjects. As can be seen from Table 5.7, the performance degradation mainly comes from pose

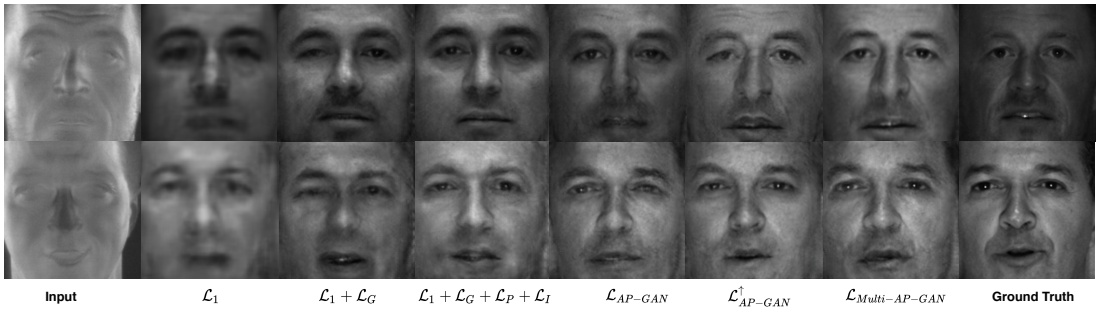


Figure 5.15: The visual results of the ablation study for different experimental settings. Given input polarimetric thermal image, synthesized results using different combination of losses and resolutions are shown successively from left to right. One intermediate synthesis results $\mathcal{L}^{\uparrow}_{AP-GAN}$, which utilizes 2-scale resolutions 128×128 and 256×256, is shown here to demonstrate the progressive improvements obtained by adding multi-scale.

variations.

5.3.2 Results on the Visible and Thermal Paired Face Database

Table 5.8 shows the performance of different methods on the Visible and Thermal Paired Face Database. Compared to the ARL Face dataset, the performance of every method is lower on this dataset. This is mainly due to the fact that this dataset is small in size and contains many facial variations. In general, the proposed method performs favorably against the previous methods.

In addition, following the analysis presented in [162], we also analyze how different variations (i.e. illumination, pose, expression, occlusion) influence the cross-spectrum matching performance of our method. As can be seen from the results in Table 5.9 illumination and pose variations are the two variations that affect the performance of our method the most. This analysis is based on the proposed method implemented with the

ground-truth visual attributes.

We also show some visual results in Fig. 5.13. It can be observed that Pix2Pix [159] and CycleGAN [165] methods generate poor quality images with many artifacts. GAN-VFS *et al.* [26] is able to synthesize better quality images. However, this method also introduces some artifacts around the eyes and mouth regions. The proposed Multi-AP-GAN method not only generates photo-realistic images but also preserves attributes on the synthesized images. We also show some images in Fig. 5.14 in which the proposed method is not able to produce good quality images. From these images we see that extreme pose, occlusion and illumination variations cause the proposed method to produce poor quality images.

5.3.3 Results on the Tufts Face Database

Table 5.10 and Fig. 5.11 show the performance of different methods on the Tufts Face Dataset [37]. Compared to the previous two datasets, this dataset is more challenging due to a large number of pose and expression variations as well as a few number of images per variation, which leads to the lower performance of every method. In general, our method outperforms the other baseline methods on this challenging dataset by improvements on 1.8 % EER and 2.4% AUC scores respectively.

Table 5.4: ARL Protocol I verification performance comparisons among the baseline methods, state-of-the-art methods, and the proposed Multi-AP-GAN method for both polarimetric thermal (Polar) and conventional thermal (S0) cases.

Method	AUC(Polar)	AUC(S0)	EER(Polar)	EER(S0)
Raw	50.35%	58.64%	48.96%	43.96%
Mahendran *et al.* [164]	58.38%	59.25%	44.56%	43.56%
Riggan *et al.* [27]	75.83%	68.52%	33.20%	34.36%
GAN-VFS *et al.* [26]	79.90%	79.30%	25.17%	27.34%
Riggan *et al.* [28]	85.43%	82.49%	21.46%	26.25%
AP-GAN [63]	$88.93\% \pm 1.54\%$	$84.16\% \pm 1.54\%$	$19.02\% \pm 1.69\%$	$23.90\% \pm 1.52\%$
AP-GAN (GT) [63]	$91.28\% \pm 1.68\%$	$86.08\% \pm 2.68\%$	$17.58\% \pm 2.36\%$	$23.13\% \pm 3.02\%$
Multi-stream GAN [34]	**96.03%**	85.74%	11.78%	23.18%
Multi-AP-GAN (ours)	$93.61\% \pm 1.46\%$	$\mathbf{90.14\% \pm 2.17\%}$	$14.24\% \pm 1.91\%$	$\mathbf{18.20\% \pm 2.65\%}$
Multi-AP-GAN (GT) (ours)	$95.29\% \pm 1.39\%$	$\mathbf{92.72\% \pm 2.03\%}$	$\mathbf{11.22\% \pm 1.89\%}$	$\mathbf{16.05\% \pm 2.15\%}$

Table 5.5: ARL Protocol II verification performance comparisons among the baseline methods and the proposed Multi-AP-GAN method for both polarimetric thermal (Polar) and conventional thermal (S0) cases.

Method	AUC (Polar)	AUC(S0)	EER(Polar)	EER(S0)
Raw	66.85%	63.66%	37.85%	40.93%
Pix2Pix [159]	$93.66\% \pm 1.07\%$	$85.09\% \pm 1.48\%$	$13.73\% \pm 1.38\%$	$23.12\% \pm 1.14\%$
Pix2PixBEGAN [159, 166]	$92.16\% \pm 1.09\%$	$83.69\% \pm 1.28\%$	$15.38\% \pm 1.45\%$	$26.22\% \pm 1.16\%$
CycleGAN [165] (supervised)	$93.11\% \pm 1.02\%$	$87.29\% \pm 1.13\%$	$15.19\% \pm 1.02\%$	$20.99\% \pm 1.19\%$
Multi-stream GAN [34]	**98.00%**	–	7.99%	–
Multi-AP-GAN (ours)	$96.55\% \pm 1.12\%$	$\mathbf{91.43\% \pm 0.93\%}$	$10.17\% \pm 1.01\%$	$\mathbf{15.86\% \pm 2.13\%}$
Multi-AP-GAN (GT) (ours)	$97.68\% \pm 0.78\%$	$\mathbf{91.88\% \pm 0.87\%}$	$\mathbf{7.69\% \pm 1.39\%}$	$\mathbf{15.29\% \pm 2.36\%}$

Table 5.6: ARL Protocol III verification performance comparisons among the baseline methods and the proposed method for both polarimetric thermal (Polar) and conventional thermal (S0) cases.

Method	AUC (Polar)	AUC(S0)	EER(Polar)	EER(S0)
Raw	73.43%	76.71%	33.56%	30.76%
Pix2Pix [159]	$86.78\% \pm 1.84\%$	$86.65\% \pm 1.48\%$	$21.92\% \pm 1.26\%$	$23.12\% \pm 1.77\%$
Pix2PixBEGAN [159, 166]	$71.29\% \pm 1.88\%$	$69.42\% \pm 1.84\%$	$33.83\% \pm 1.68\%$	$36.88\% \pm 1.76\%$
CycleGAN [165] (supervised)	$86.77\% \pm 1.77\%$	$81.80\% \pm 1.67\%$	$21.48\% \pm 1.11\%$	$25.86\% \pm 1.36\%$
GAN-VFS [3] [26]	$90.20\% \pm 1.85\%$	$87.10\% \pm 1.52\%$	$18.53\% \pm 1.21\%$	$20.22\% \pm 1.92\%$
Multi-AP-GAN (ours)	$\mathbf{92.29\% \pm 1.48\%}$	$\mathbf{88.49\% \pm 1.87\%}$	$\mathbf{16.26\% \pm 1.12\%}$	$\mathbf{19.25\% \pm 1.62\%}$
Multi-AP-GAN (GT) (ours)	$\mathbf{93.72\% \pm 1.08\%}$	$\mathbf{90.99\% \pm 1.13\%}$	$\mathbf{14.75\% \pm 1.36\%}$	$\mathbf{17.81\% \pm 1.63\%}$

Table 5.7: Protocol III verification performance with respect to different variations.

Variations	AUC (Polar)	AUC(S0)	EER(Polar)	EER(S0)
Neutral	$96.77\% \pm 1.25\%$	$94.69\% \pm 1.17\%$	$12.50\% \pm 2.09\%$	$13.38\% \pm 1.48\%$
Expression	$96.77\% \pm 1.91\%$	$92.38\% \pm 1.40\%$	$10.05\% \pm 2.02\%$	$15.18\% \pm 1.58\%$
Pose	$86.62\% \pm 2.39\%$	$82.35\% \pm 2.54\%$	$22.45\% \pm 1.84\%$	$25.76\% \pm 1.95\%$
Average	$93.72\% \pm 1.08\%$	$90.99\% \pm 1.13\%$	$14.75\% \pm 1.36\%$	$17.81\% \pm 1.63\%$

Table 5.8: Visible and Thermal Paired Face Database verification performance comparisons among the baseline methods and the proposed method for the conventional thermal case.

Method	AUC	EER
Raw	69.54%	35.39%
Pix2Pix [159]	$78.66\% \pm 1.48\%$	$28.39\% \pm 1.14\%$
Pix2PixBEGAN [159, 166]	$73.69\% \pm 1.82\%$	$34.22\% \pm 1.61\%$
CycleGAN [165] (supervised)	$80.24\% \pm 1.31\%$	$26.72\% \pm 1.39\%$
GAN-VFS [3] [26]	$80.44\% \pm 1.03\%$	$26.33\% \pm 1.19\%$
CRN + CL [3] [68, 69]	$81.25\% \pm 1.01\%$	$26.01\% \pm 1.23\%$
Multi-AP-GAN (ours)	$\mathbf{81.73\% \pm 0.93\%}$	$\mathbf{25.68\% \pm 1.56\%}$
Multi-AP-GAN (GT) (ours)	$\mathbf{82.68\% \pm 0.87\%}$	$\mathbf{23.16\% \pm 0.98\%}$

Table 5.9: Verification performance with respect to different variations on the Visible and Thermal Paired Face Database.

Variations	AUC	EER
Illumination	$73.35\% \pm 0.25\%$	$32.60\% \pm 0.43\%$
Expression	$97.25\% \pm 0.68\%$	$7.45\% \pm 1.74\%$
Pose	$78.25\% \pm 1.03\%$	$28.75\% \pm 0.93\%$
Occlusion	$83.98\% \pm 1.33\%$	$24.02\% \pm 1.06\%$
Average	$82.68\% \pm 0.87\%$	$23.16\% \pm 0.98\%$

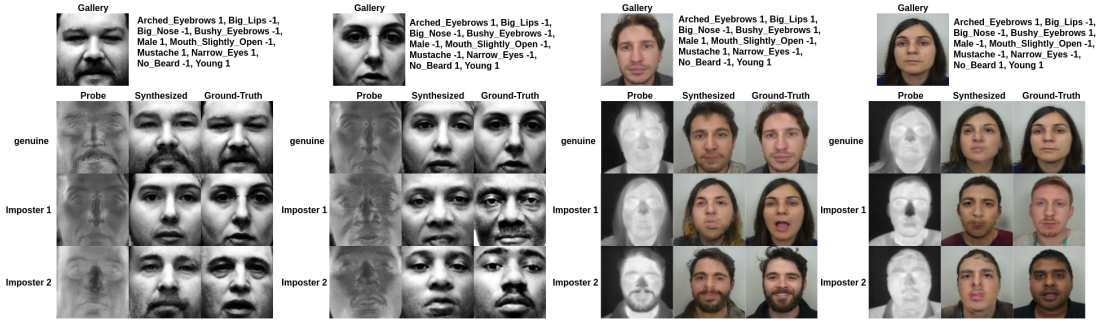


Figure 5.16: Analysis of attributes on synthesis. We show the synthesis samples from either conventional or polarimetric thermal images on both datasets. Given probe (thermal) images and estimated attributes from the gallery (visible) image, our proposed method can generates attribute preserving (visible) images.

Table 5.10: The Tufts Face Database [37] verification performance comparisons among the baseline methods and the proposed method.

Method	AUC	EER
Raw	66.73%	38.13%
Pix2Pix [159]	$69.73\% \pm 0.92\%$	$35.83\% \pm 0.59\%$
Pix2PixBEGAN [159, 166]	$68.89\% \pm 0.51\%$	$36.88\% \pm 0.43\%$
CycleGAN [165] (supervised)	$71.93\% \pm 1.94\%$	$34.16\% \pm 1.70\%$
GAN-VFS [3] [26]	$73.78\% \pm 0.46\%$	$32.32\% \pm 0.53\%$
CRN + CL [3] [68, 69]	$74.90\% \pm 0.56\%$	$31.71\% \pm 0.54\%$
Multi-AP-GAN (ours)	$\mathbf{75.86\% \pm 0.88\%}$	$\mathbf{31.14\% \pm 0.74\%}$
Multi-AP-GAN (GT) (ours)	$\mathbf{77.38\% \pm 0.98\%}$	$\mathbf{29.94\% \pm 0.79\%}$

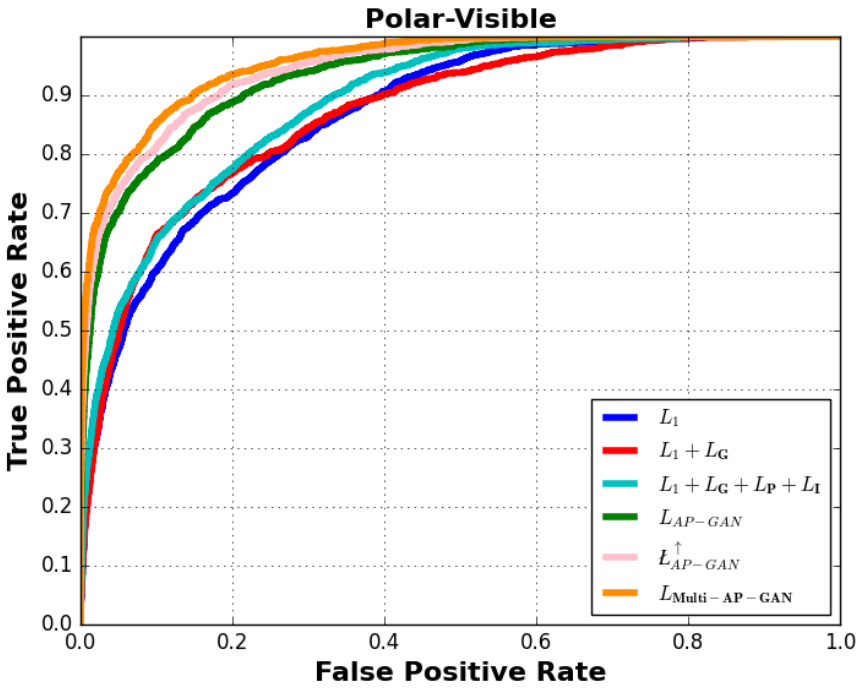


Figure 5.17: The ROC curves corresponding to the ablation study.

5.3.4 Ablation Study

In order to demonstrate the effectiveness of different modules in the proposed method, we conduct the following ablation study using the Polarimetric thermal modality in the ARL dataset on Protocol I:

1. Polar to Visible estimation with only $\mathcal{L}_1$ (as defined in Eq. (5.5))
2. Polar to Visible estimation with $\mathcal{L}_1$ and $\mathcal{L}_G$ (as defined in Eq. (5.2))
3. Polar to Visible estimation with $\mathcal{L}_1$, $\mathcal{L}_G$, perceptual loss $\mathcal{L}_P$ and identity loss $\mathcal{L}_I$, which are defined as in Eq. (5.4).

4. Polar to Visible estimation with all the losses as defined in Eq. (5.3), by utilizing various solution scales: $\mathcal{L}_{AP-GAN}$ (256^2), $\mathcal{L}^{\uparrow}_{AP-GAN}$ $(128^2, 256^2)$, $\mathcal{L}_{Multi-AP-GAN}$ $(64^2, 128^2, 256^2)$ respectively.

Fig. 5.17 shows the ROC curves corresponding to each experimental setting. From this figure, we can observe that using all the losses together as $\mathcal{L}_{Multi-AP-GAN}$ can obtain the best performance. Compared to the results between $\mathcal{L}_1$ and $\mathcal{L}_1 + \mathcal{L}_G$, we can observe the enhancement provided by adding the adversarial loss. Compared with the results between $\mathcal{L}_1 + \mathcal{L}_G$ and $\mathcal{L}_1 + \mathcal{L}_G + \mathcal{L}_P + \mathcal{L}_I$, we can observe the improvements obtained by adding the perceptual and identity loses. On the other hand, one can clearly see the significance of fusing the semantic attribute information with the image feature in the latent space by comparing the results between $\mathcal{L}_1 + \mathcal{L}_G + \mathcal{L}_P + \mathcal{L}_I$ and $\mathcal{L}_{AP-GAN}$. Additionally, looking at the comparison with $\mathcal{L}_{AP-GAN}$, $\mathcal{L}^{\uparrow}_{AP-GAN}$ and $\mathcal{L}_{Multi-AP-GAN}$, one can see the successive improvements by leveraging the multi-scale information.

Besides the ROC curves, we also show the visual results for each experimental setting in Fig. 5.15. Given the input Polar image, the synthesized results from different experimental settings are shown in Fig. 5.15. It can be observed that $\mathcal{L}_1$ captures the low-frequency features of images very well. $\mathcal{L}_1 + \mathcal{L}_G$ can capture both low-frequency and high-frequency features in the image. However, it adversely introduced distortions and artifacts in the synthesized image. In addition, optimizing $\mathcal{L}_P + \mathcal{L}_I$ suppresses these distortions to some extent. Finally, fusing attributes into the loss on with leveraging multi-scale resolution (i.e. $\mathcal{L}_{Multi-AP-GAN}$) can not only improving the performance but also preserves facial

attributes. In our study, we do not see significant more improvement by utilizing more than 3-scale resolutions.

In addition, we analyze the effect of attributes on the synthesized images in Figure 5.16. In particular, given the input gallery image, we examine how attributes help in synthesizing a visible image from a thermal probe image. If the probe image and the input gallery image share the same identity then Multi-AP-GAN is able to generate attribute preserving visible image. On the other hand, if the probe image's identity is different from that of the gallery image then the proposed method is not able to synthesize identity preserving visible face. However, the attributes are still preserved on the synthesized image. This analysis further demonstrates that the proposed Multi-AP-GAN method learns the cross-spectral (thermal-to-visible) translation mapping exactly guided by the visual attributes.

5.4 Discussion

The proposed Multi-AP-GAN approach generates better quality visible images and as a result obtains improved cross-modal verification performance compard to previous GAN-based approaches. This can be contributed to the fact that Multi-AP-GAN uses a better generator which is guided by visual attributes. The multi-scale generator mitigates the receptive-field limitation of the convolutional operation by leveraging the features corresponding to images at multiple scales. In addition, visual attributes provide complementary

[3] results are obtained after re-implementation due to the limited code availability.
[1] features are extracted: https://github.com/TreB1eN/InsightFace_Pytorch

semantic information for better synthesis. GAN-based methods such as GAN-VFS [26], Multi-stream GAN [34] and Pix2Pix [159] are single-scale generators and do not exploit such facial semantic information during synthesis.

Though our method performs reasonably well on three datasets, there are some limitations which we hope to overcome in our future work. Our model requires paired thermal and visible face images for training, which is laborious and expensive. Hence, an unsupervised synthesis method that does not require paired data is needed. Another limitation of our approach is that it does not work well on extreme pose variations. We are currently developing a new method that can deal with this pose issue in heterogeneous face recognition. We also plan to further investigate the impact of metabolic and physiologic variability in thermal facial signatures on synthesis and subsequent recognition performance.

5.5 Summary

We propose a novel Attribute Preserving Generative Adversarial Network (Multi-AP-GAN) structure for thermal-to-visible face verification via synthesizing photo-realistic visible face images from the corresponding thermal (polarimetric or conventional) images with extracted attributes. Rather than use only image-level information for synthesis and verification, we take a different approach in which semantic facial attribute information is also fused during training and testing. Quantitative and visual experiments evaluated on a real thermal-visible dataset demonstrate that the proposed method achieves state-of-

the-art performance compared with other existing methods. In addition, an ablation study is developed to demonstrate the improvements obtained by different combination of loss functions.

Chapter 6

Polarimetric Thermal to Visible Face Verification via Self-Attention Guided Synthesis

Recognizing faces in low-light/night-time with that in normal (visible) conditions is a very difficult problem. Various thermal imaging modalities have been introduced in the literature to deal with this problem. The infrared spectrum can be divided into a reflection dominated region consisting of the near infrared (NIR) and shortwave infrared (SWIR) bands, and an emission dominated thermal region consisting of the midwave infrared (MWIR) and longwave infrared (LWIR) bands [167]. It has been shown that polarimetric thermal imaging captures additional geometric and textural facial details compared to conventional thermal imaging [140]. Hence, the polarization-state information has been

used to improve the performance of cross-spectrum face recognition [26–28, 63, 140, 152].

In polarimetric thermal to visible face verification, given a pair of visible and polarimetric thermal images, the goal is to determine whether these images correspond to the same person. The large domain discrepancy between these images makes the cross-spectrum matching problem very challenging. Various methods have been proposed in the literature for cross-spectrum matching [2, 26–28, 140, 152–155, 168, 169]. These approaches either attempt to synthesize visible faces from thermal faces or extract robust features from these modalities for cross-modal matching.

In this work, we take a different approach to the problem of thermal to visible matching by exploring the complementary information of different modalities. Figure 6.1 gives an overview of the proposed approach. Given a thermal-visible pair $(\mathbf{x}_t, \mathbf{x}_v)$, these images are first transformed into their spectrum counterparts using two trained generators as $\hat{\mathbf{x}}_v = G_{t\to v}(\mathbf{x}_t)$, $\hat{\mathbf{x}}_t = G_{v\to t}(\mathbf{x}_v)$. Then a feature extractor network $Feat$, in particular the VGG-Face model [145], is used to extract features $f_{\mathbf{x}_t} = Feat(\mathbf{x}_t)$, $f_{\hat{\mathbf{x}}_v} = Feat(\hat{\mathbf{x}}_v)$, $f_{\mathbf{x}_v} = Feat(\mathbf{x}_v)$, and $f_{\hat{\mathbf{x}}_t} = Feat(\hat{\mathbf{x}}_t)$. These features are then fused to generate the gallery template $g_{\mathbf{x}_v} = (f_{\mathbf{x}_v} + f_{\hat{\mathbf{x}}_t})/2$ and the probe template $g_{\mathbf{x}_t} = (f_{\mathbf{x}_t} + f_{\hat{\mathbf{x}}_v})/2$. Finally, the cosine similarity score between these feature templates is calculated for verification.

6.1 Proposed Method

In this chapter, we discuss details of the proposed self-attention guided synthesis method. In particular, we discuss the proposed generator and the discriminator networks as well as the loss functions used to train the network. The overall framework is shown in Figure 6.2. Given an input image from one modality (thermal as shown), it is first synthesized into the other modality (i.e. visible) using the proposed self-attention module-based generator. Then another generator with similar architecture is used to synthesize it back from the visible domain to the original thermal domain. In order to achieve the reconstruction back to original modality, these generators are trained using the cycle-consistency loss [120, 165]. In order to minimize the domain gap between the fake (i.e. synthesized) and real images, a patch-based pixel GAN loss is also introduced [159]. Furthermore, the semantic and identity information are captured by minimizing the perceptual and identity loss [106], respectively.

6.1.1 Generator

An encoder-decoder type of generator which is inspired by the residual network (He *et al.* [103]) and SAGAN (Han *et al.* [170]) is adopted in this work. In order to prevent the vanishing gradient problem, the residual block is implemented after a sequence of convolutional layers. For each residual block shown in Figure 6.3, it consists of two convolutional layers followed by batch-normalization and relu layers. In order to involve the facial

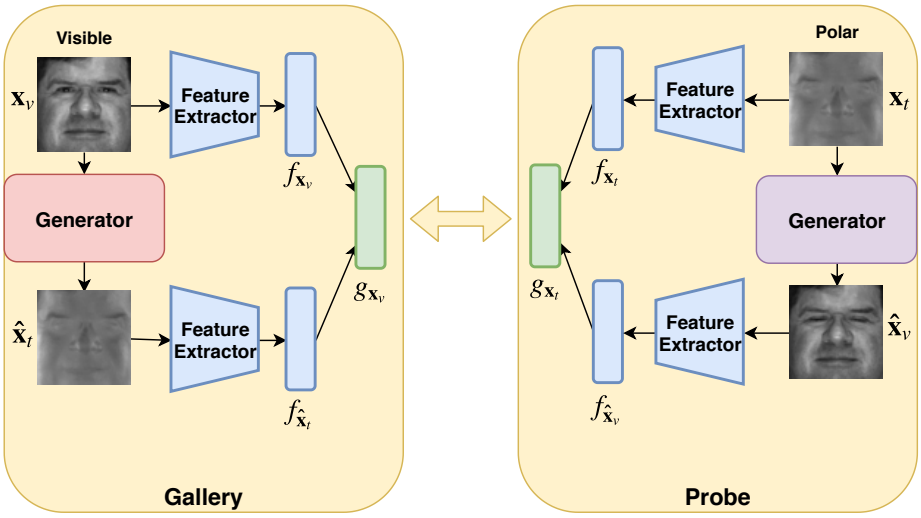


Figure 6.1: An overview of the proposed cross-modal face verification method. Given a visible gallery image $\mathbf{x}_v$, a generator network is used to synthesize the corresponding thermal image $\hat{\mathbf{x}}_t$. Similarly, given a polarimetric thermal probe image $\mathbf{x}_t$, a different generator network is used to synthesize the corresponding visible image $\hat{\mathbf{x}}_v$. Pre-trained CNNs are used to extract features from the original and the synthesized images. These features are then fused to generate the gallery template $g_{\mathbf{x}_v}$ and the probe template $g_{\mathbf{x}_t}$. Finally, the cosine similarity score between these feature templates is calculated for verification.

long-range dependency information, we adopt the self-attention module into the generator. Self-attention module was proposed by Han *et al.*in SAGAN [170] which allows attention-driven, long-range dependency modeling for general image generation tasks. In our work, the self-attention module is inserted right before the last convolutional layer of the generator. The self-attention module, shown as in Figure 6.3, consists of two components: feature maps and attention maps. The feature maps are generated by a 1×1 convolutional layer working on the input features. The attention maps are generated by the elementwise multiplication of two 1×1 convolutional features followed by the softmax function. Finally, this module outputs the elementwise multiplication of feature maps and attention maps.

The self-attention module-based generator architecture is shown in Figure 6.3. This

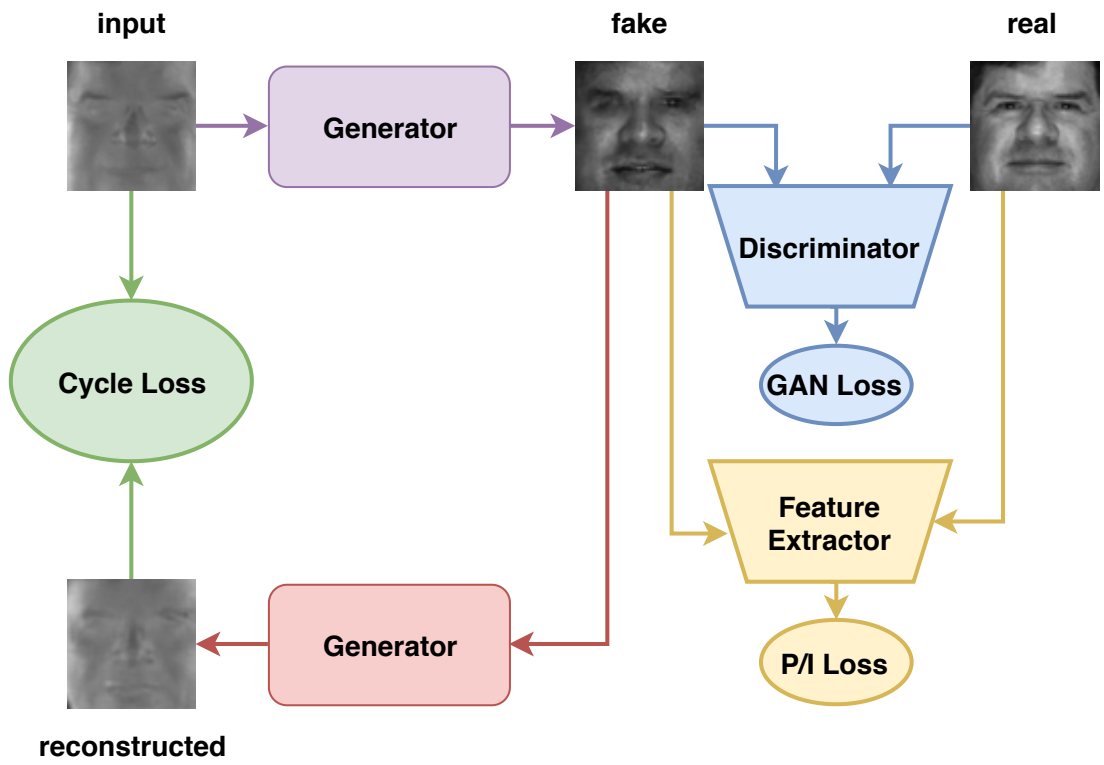


Figure 6.2: Self-attention guided synthesis of visible images from polarimetric thermal input. In order to minimize the domain gap between different modalities, the input thermal/visible images are directly mapped into the visible/thermal modality. In order to obtain the image level style, the pixel GAN loss (blue) and cycle consistancy loss (green) are introduced. The feature-level semantic information is captured by the identity and perceptual losses (yellow). Similar architecture can also be used for synthesizing thermal images from visible images.

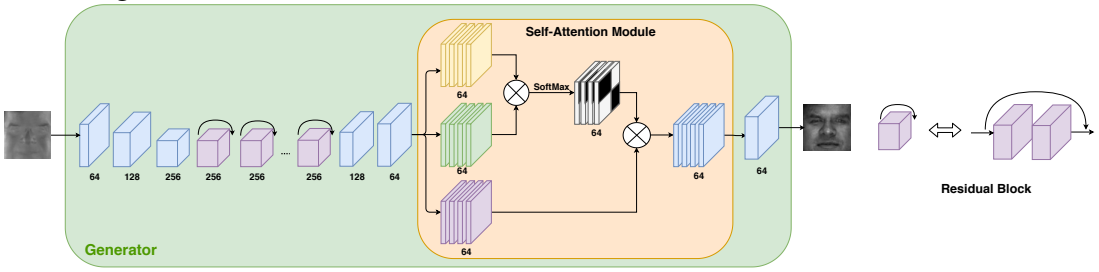


Figure 6.3: The proposed self-attention module-based generator architecture.

generator architecture is consists of the following components:

CBR(64)-CBR(128)-CBR(256)-Res(256)-Res(256)-Res(256)-Res(256)-Res(256)-DBL(128)-

DBL(64)-SA(64)-CT(3),

where C stands for the convolutional layer (stride 2, kernel-size 4, and padding size 1). B and R stand for batch-normalization layer and relu layer, respectively. Res is the residual block [103], D denotes the deconvolutional layer (stride 2, kernel-size 4 and padding size 1), L is the leaky relu layer, SA is the self-attention module, and T is the tanh function layer. The numbers inside the parenthesis denote the number of channels corresponding to the output feature maps.

6.1.2 Discriminator

Motivated by pixel GAN [159], a patch-based discriminator is leveraged in the proposed method and is trained iteratively with the generator. In addition, in order to improve the stability of training, we adopt the spectral normalization to the discriminator [158]. Compared to the other normalization techniques, spectral normalization does not require extra hyper-parameter tuning and has a relatively small the computational cost. Similarly, in order to capture the long-range dependency information, a self-attention module is added before the last convolutional layer in the discriminator. The discriminator, as shown in Figure 6.4, consists of the following components:

CLSn(64)-CLSn(128)-CLSn(256)-CLSn(512)-CLS(n512)-SA(512)-CS(1),

where Sn stands for the spectral normalization layer. C, L, SA, S stand for the convolutional layer, leaky relu layer, self-attention module and sigmoid function, respectively. The numbers inside the parenthesis denote the number of channels corresponding to the output

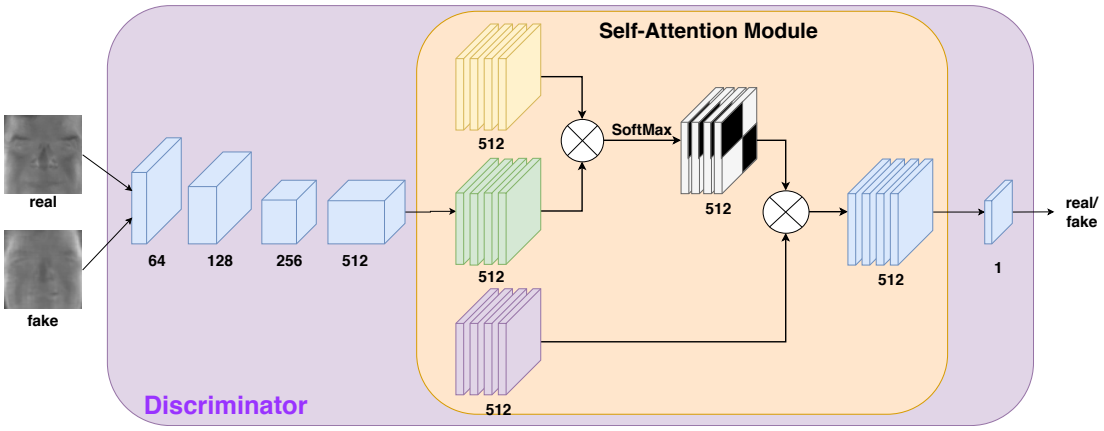
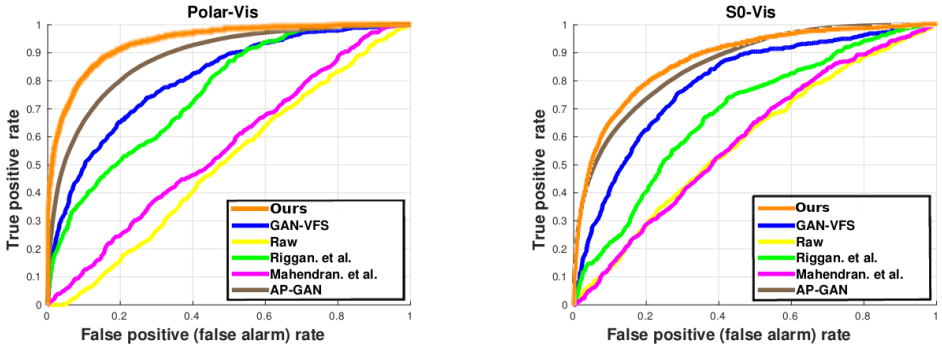


Figure 6.4: The architecture of the proposed discriminator.



(a) (b)

Figure 6.5: The ROC curve comparison on Protocol I with several state-of-the-art methods: GAN-VFS [26], Riggan *et al.* [27] Mahendran *et al.* [164], AP-GAN [63]. (a) The performance on Polar-Visible verification. (b) The performance on S0-Visible verification.

feature maps.

6.1.3 Objective Function

Given a set of thermal images $\mathbf{X}_t = \{x_t^i\}_{i=1}^N$ and another set of visible images $\mathbf{X}_v = \{x_v^i\}_{i=1}^N$, the generator and discriminator networks are optimized iteratively by minimizing

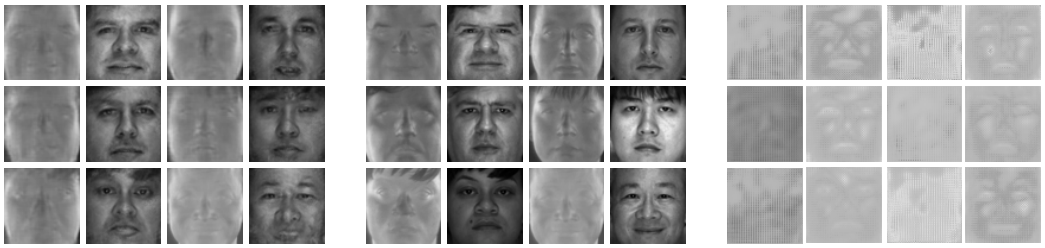


Figure 6.6: (a) Sample synthesized results on both visible and thermal modalities. (b) Reference images. (c) The learned self-attention feature maps. Images corresponding to different modality are shown in different columns.

the following loss functions

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{GAN}(G_{t\to v}, D_v, \mathbf{X}_t, \mathbf{X}_v) + \mathcal{L}_{GAN}(G_{v\to t}, D_t, \mathbf{X}_v, \mathbf{X}_t)\\ +\lambda_P\mathcal{L}_P(G_{t\to v}, \mathbf{X}_t, \mathbf{X}_v) + \lambda_P\mathcal{L}_P(G_{v\to t}, \mathbf{X}_v, \mathbf{X}_t)\\ +\lambda_I\mathcal{L}_I(G_{t\to v}, \mathbf{X}_t, \mathbf{X}_v) + \lambda_I\mathcal{L}_I(G_{v\to t}, \mathbf{X}_v, \mathbf{X}_t)\\ +\lambda_1\mathcal{L}_1(G_{t\to v}, \mathbf{X}_t, \mathbf{X}_v) + \lambda_1\mathcal{L}_1(G_{v\to t}, \mathbf{X}_v, \mathbf{X}_t)\\ +\mathcal{L}_{cycle}(G_{t\to v}, G_{v\to t}, \mathbf{X}_t, \mathbf{X}_v),\end{aligned}$$

where $\mathcal{L}_{GAN}(G_{t\to v}, D_v, \mathbf{X}_t, \mathbf{X}_v), \mathcal{L}_{GAN}(G_{v\to t}, D_t, \mathbf{X}_v, \mathbf{X}_t)$ are the adversarial losses for two generators - one for synthesizing visible from thermal ($G_{t\to v}$) and the other for synthesizing thermal from visible ($G_{v\to t}$). Similarly, $\mathcal{L}_P$ is the perceptual loss, $\mathcal{L}_I$ is the identity loss, $\mathcal{L}_1$ is the loss based on the L1-norm between the target and the synthesized image, and $\lambda_P, \lambda_I, \lambda_1$ are the weights for perceptual loss, identity loss and L1 loss, respectively.

6.1.3.1 Adversarial Loss

Similar to the Cycle-GAN work [165], there are two kinds adversarial losses. One $\mathcal{L}_{GAN}(G_{t\to v}, D_v, \mathbf{X}_t, \mathbf{X}_v)$ for synthesizing visible image from thermal image and the other $\mathcal{L}_{GAN}(G_{v\to t}, D_t, \mathbf{X}_v, \mathbf{X}_t)$ for synthesizing thermal image from visible image. Both are defined as follows:

$$\begin{aligned}\mathcal{L}_{GAN}(G_{t\to v}, D_v, \mathbf{X}_t, \mathbf{X}_v) &= \mathbb{E}_{\mathbf{x}_v\sim\mathbf{X}_v}[\log D_v(\mathbf{x}_v)],\\ &+\mathbb{E}_{\mathbf{x}_t\sim\mathbf{X}_t}[\log(1 - D_v(G_{t\to v}(\mathbf{x}_t))]\\ \mathcal{L}_{GAN}(G_{v\to t}, D_t, \mathbf{X}_v, \mathbf{X}_t) &= \mathbb{E}_{\mathbf{x}_t\sim\mathbf{X}_t}[\log D_t(\mathbf{x}_t)]\\ &+\mathbb{E}_{\mathbf{x}_v\sim\mathbf{X}_v}[\log(1 - D_t(G_{v\to t}(\mathbf{x}_v))],\end{aligned} \tag{6.1}$$

where D_v and D_t are discriminators for visible and thermal modality, respectively. In addition, $G_{v\to t}$ and $G_{v\to t}$ are two generators for synthesizing thermal image from visible and synthesizing visible image from thermal, respectively.

6.1.3.2 Cycle-Consistency Loss

A cycle-consistency constraint is also imposed in our approach [120, 165] (see Figure 6.2 green portion). Taking thermal to visible synthesis as an example, we introduce one mapping from thermal to visible $G_{t\to v}$ and train it according to the same GAN loss $\mathcal{L}_{GAN}(G_{t\to v}, D_v, \mathbf{X}_t, \mathbf{X}_v)$. We then require another mapping from thermal to visible and back to thermal which reproduces the original sample, thereby enforcing cycle-consistency.

In other words, we want $G_{v\to t}(G_{t\to v}(\mathbf{x}_t)) \sim \mathbf{x}_t$ and $G_{t\to v}(G_{v\to t}(\mathbf{x}_v)) \sim \mathbf{x}_v$. This is done by imposing an L_1 penalty on the reconstruction error, which is referred to as the cycle-consistency loss. It is defined as follows:

$$\begin{aligned} &\mathcal{L}_{cycle}(G_{t\to v}, G_{v\to t}, \mathbf{X}_t, \mathbf{X}_v) = \\ &\mathbb{E}_{\mathbf{x}_t\sim\mathbf{X}_t}[\|G_{v\to t}(G_{t\to v}(\mathbf{x}_t)) - \mathbf{x}_t\|_1] \\ &+\mathbb{E}_{\mathbf{x}_v\sim\mathbf{X}_v}[\|G_{t\to v}(G_{v\to t}(\mathbf{x}_v)) - \mathbf{x}_v\|_1]. \end{aligned} \tag{6.2}$$

6.1.3.3 Perceptual, Identity and L1 Loss Functions

These loss functions can be implemented when we have supervised pairwise data $\{(\mathbf{x}_t^i, \mathbf{x}_v^i)\}_{i=1}^N$, where $\mathbf{x}_t^i \in \mathbf{X}_t$ and $\mathbf{x}_v^i \in \mathbf{X_v}$, during training. The L1 loss is defined as below:

$$\begin{aligned} \mathcal{L}_1(G_{t\to v}, \mathbf{x}_t^i, \mathbf{x}_v^i) &= \|G_{t\to v}(\mathbf{x}_t^i) - \mathbf{x}_v^i\|_1 \\ \mathcal{L}_1(G_{v\to t}, \mathbf{x}_v^i, \mathbf{x}_t^i) &= \|G_{v\to t}(\mathbf{x}_v^i) - \mathbf{x}_t^i\|_1. \end{aligned} \tag{6.3}$$

In order to minimize the perceptual and identity information [106, 160], we implement the perceptual and identity loss functions as follows

$$\mathcal{L}_p(G_{t\to v}, F_p, \mathbf{x}_t^i, \mathbf{x}_v^i) = [\|F_p(G_{t\to v}(\mathbf{x}_t^i)) - F_p(\mathbf{x}_v^i)\|_1]$$

$$\mathcal{L}_p(G_{v\to t}, F_p, \mathbf{x}_v^i, \mathbf{x}_t^i) = [\|F_p(G_{v\to t}(\mathbf{x}_v^i)) - F_p(\mathbf{x}_t^i)\|_1]$$

$$\mathcal{L}_I(G_{t\to v}, F_I, \mathbf{x}_t^i, \mathbf{x}_v^i) = [\|F_I(G_{t\to v}(\mathbf{x}_t^i)) - F_I(\mathbf{x}_v^i)\|_1]$$

$$\mathcal{L}_I(G_{v\to t}, F_I, \mathbf{x}_v^i, \mathbf{x}_t^i) = [\|F_I(G_{v\to t}(\mathbf{x}_v^i)) - F_I(\mathbf{x}_t^i)\|_1],$$

where F_I and F_p are two off-the-shelf pretrained networks for extracting features. Since deeper features in hierarchical deep networks capture more semantic information, the output features $conv2_2$ and $conv4_2$ from the VGGFace pretrained network are used in the perceptual and the identity losses, respectively.

Note that if we omit the perceptual, identity and L1 loss functions which require pairwise supervised data, then one can also implement the proposed framework in completely unsupervised fashion. In other words, the proposed framework is also applicable to the case where the paired data are not available during training.

Table 6.1: Protocol I Verification performance comparisons among the baseline methods and the proposed method for both polarimetric thermal (Polar) and conventional thermal (S0) cases.

Method	AUC (Polar)	AUC(S0)	EER(Polar)	EER(S0)
Raw	50.35%	58.64%	48.96%	43.96%
Mahendran *et al.* [164]	58.38%	59.25%	44.56%	43.56%
Riggan *et al.* [27]	75.83%	68.52%	33.20%	34.36%
GAN-VFS [26]	79.90%	79.30%	25.17%	27.34%
Riggan *et al.* [28]	85.42%	82.49%	21.46%	26.25%
AP-GAN [63]	$88.93\% \pm 1.54\%$	$84.16\% \pm 1.54\%$	$19.02\% \pm 1.69\%$	$23.90\% \pm 1.52\%$
Multi-stream GAN [34]	$\mathbf{96.03\%}$	85.74%	$\mathbf{11.78\%}$	23.18%
Ours	$93.68\% \pm 0.97\%$	$\mathbf{89.20\% \pm 1.56\%}$	$13.46\% \pm 1.92\%$	$\mathbf{18.77\% \pm 1.36\%}$

Table 6.2: Protocol II Verification performance comparisons among the baseline methods and the proposed method for both polarimetric thermal (Polar) and conventional thermal (S0) cases.

Method	AUC (Polar)	AUC(S0)	EER(Polar)	EER(S0)
Raw	66.85%	63.66%	37.85%	40.93%
CycleGAN [165](unsupervised)	$76.09\% \pm 1.49\%$	$74.17\% \pm 1.34\%$	$32.28\% \pm 1.68\%$	$33.04\% \pm 1.39\%$
ours(unsupervised [1])	$86.92\% \pm 1.42\%$	$80.02\% \pm 1.16\%$	$21.51\% \pm 1.24\%$	$28.09\% \pm 1.04\%$
Pix2Pix [159]	$93.66\% \pm 1.07\%$	$85.09\% \pm 1.48\%$	$13.73\% \pm 1.38\%$	$23.12\% \pm 1.14\%$
Pix2PixBEGAN [159, 166]	$92.16\% \pm 1.09\%$	$83.69\% \pm 1.28\%$	$15.38\% \pm 1.45\%$	$26.22\% \pm 1.16\%$
CycleGAN [165] (supervised)	$93.11\% \pm 1.02\%$	$87.29\% \pm 1.13\%$	$15.19\% \pm 1.02\%$	$20.99\% \pm 1.19\%$
Multi-stream GAN [34]	**98.00%**	–	**7.99%**	–
Ours	$96.41\% \pm 1.02\%$	$\mathbf{91.49\% \pm 2.25\%}$	$10.02\% \pm 0.03\%$	$\mathbf{15.45\% \pm 2.31\%}$

6.2 Experimental Results

The proposed method is evaluated on the ARL Multimodal Face Database [140] which consists of polarimetric (i.e. Stokes image) and visible images from Volume I [140] and II [34]. The Volume I data consists of images corresponding to 60 subjects. On the other hand, the Volume II data consists of images from 51 subjects (81 subjects in total). Similar to [26, 27, 34], we evaluate the proposed method based on two protocols. For Protocol I, images corresponding to Range 1 from 30 subjects are used for training. The remaining 30 subjects' data are used for evaluation. For Protocol II, all images from 81 subjects are used for running experiments. Specifically, all images from Volume I and 25 subjects' images from Volume II are used for training, the remaining 26 subjects' images from Volume II are used for evaluation. We repeat this process 5 times and report the average results.

We evaluate the face verification performance of proposed method and compare it with several recent works [26, 28, 34, 159, 165]. Moreover, the performance is evaluated based

on the FC-7 layer of the pretrained VGG-Face model [145] using the receiver operating characteristic (ROC) curve, Area Under the Curve (AUC) and Equal Error Rate (EER) measures. To summarize, the proposed method is evaluated on the following four experiments:

a) Conventional thermal (S0) to Visible (Vis) on Protocol I.

b) Polarimetric thermal (Polar) to Visible (Vis) on Protocol I.

c) Conventional thermal (S0) to Visible (Vis) on Protocol II.

d) Polarimetric thermal (Polar) to Visible (Vis) on Protocol II.

6.2.1 Implementation

In addition to the standard preprocessing as discussed in [140], two more preprocessing steps are used in the proposed method. First, the faces in the visible domain are detected by MTCNN [163]. Then, a standard central crop method is used to crop the registered faces. Since the MTCNN is implementable on the visible images only, we use the same detected rectangle coordination to crop the S0, S1, S2 images. After preprocessing, all the images are scaled to be 224×224 and are saved as 16-bit PNG files.

The entire network is trained in Pytorch on a single Nvidia Titan-X GPU. The L1, perceptual and identity loss parameters are chosen as λ_1=10, λ_p = 2, λ_I = 0.2 respectively by a grid search. The ADAM [110] is implemented as the optimization algorithm with parameter betas = (0.5,0.999) and batch size is chosen as 8. The total epochs are 200 for Protocol I and 100 for Protocol II. For the first half epochs, we fix the learning rate

as $lr = 0.0002$ and for the remaining epochs, the learning rate was decreased by 1/100 (Protocol I) and 1/50 (Protocol II) after each epoch.

Once the generators are trained, they could be implemented on the given probe and gallery images as shown in Figure 6.1.

6.2.2 Comparison with state-of-the-art Methods

Regarding Protocol I, we evaluate and compare the performance of the proposed method with recent state-of-the-art methods [26–28, 34, 63, 164]. Figure 6.5 shows the evaluation performance for two different experimental settings, S0 (representing conventional thermal) and Polar separately. As can be seen from Figure 6.6, compared with the other state-of-the-art methods, the proposed method performs better and comparably to [34]. In addition, it can be observed that the performance corresponding to the Polar modality is always better than the S0 modality, which demonstrates the advantage of using the polarimetric thermal images than the conventional thermal images. The quantitative comparisons are shown in Table 6.1, and also demonstrate the effectiveness of the proposed method. Furthermore, Figure 6.6 shows some synthesized images. As can be seen from this figure, the facial attributes and the identity information is preserved well. Furthermore, from Figure 6.6(c) we see that the learned self-attention maps corresponding to both visible and thermal images are always located on the facial attributes regions such as mouth, eyes, and nose. As a result, the proposed self-attention guided GAN is able to capture meaningful information from both modalities for synthesis.

Table 6.2 compares the performance of several state-of-the-art image synthesis on Protocol II. These include multi-stream GAN [34], Pix2Pix [159], CycleGAN [165], and Pix2Pix-BEGAN [166]. Note that most prior works have not reported their results on Protocol II as it is based on a new extended dataset that was only recently made publicly available. Similar to Protocol I, the experiments are evaluated on two different settings - S0 and Polar separately. As can be seen from Table 6.2, the proposed method performs comparably to the most recent state-of-the art image synthesize algorithms. In this table, we also report the unsupervised performance of different methods. As expected, the supervised results outperform the unsupervised results with a large margin. Furthermore, the proposed method in unsupervised setting performs better than the other compared method. This clearly shows the significance of using self-attention module in our framework.

Note that a GAN-based multi-stream fusion method recently proposed in [34] is a supervised method that is specifically designed for the polarimetric data. The generator network consists of a multi-stream feature-level fusion encoder-decoder network. As a result, the performance of [34] is slightly better than our method on the polar modality. On the other hand, our method outperforms [34] by a large margin when only the S0 modality is used as the input. Our method can be viewed as a generic heterogeneous face recognition method. The performance of our method can be improved by using more sophisticated generators and feature extractors.

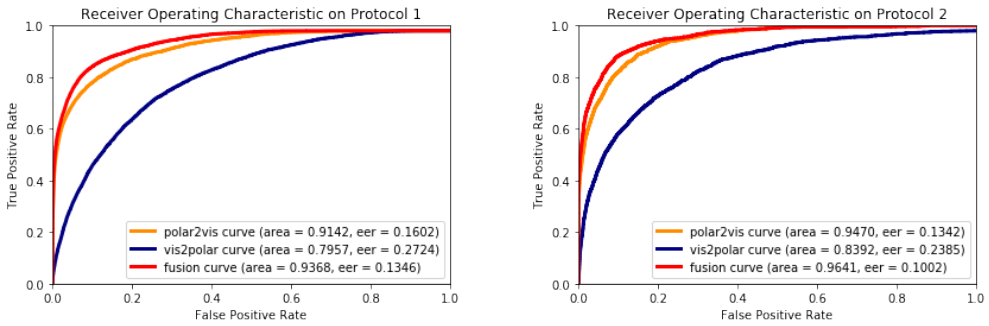


Figure 6.7: The ROC curves corresponding to the proposed fusion method as well as individual modalities.

6.2.3 Ablation Study Regarding Fusion

In this part we analyze the effectiveness of using fusion features in our method. In this ablation study, given polar (thermal) images $\mathbf{X}_t$ and visible images $\mathbf{X}_v$, we implement the following three experiments:

polar2vis: generate the visible images from the polar $\hat{\mathbf{X}}_v = G_{t\to v}(\mathbf{X}_t)$, then verify based on the features from $(\hat{\mathbf{X}}_v, \mathbf{X}_v)$.

vis2polar: generate the polar images from the visible $\hat{\mathbf{X}}_t = G_{v\to t}(\mathbf{X}_v)$, then verify based on the feature from $(\hat{\mathbf{X}}_t, \mathbf{X}_t)$.

fusion: generate the visible images from the polar $\hat{\mathbf{X}}_v = G_{t\to v}(\mathbf{X}_t)$ and the polar images from the visible $\hat{\mathbf{X}}_t = G_{v\to t}(\mathbf{X}_v)$, then verify the images based on the features from $((\hat{\mathbf{X}}_t + \mathbf{X}_v)/2, (\mathbf{X}_t + \hat{\mathbf{X}}_v)/2)$.

This experiment will clearly show the significance of generating templates by fusing two features. The ablation study is evaluated on both Protocol I and Protocol II and the results are shown in Figure 6.7. Compared to the unimodal results, the fusion method significantly improves the performance on both protocols. Also, the visible modality out-

performs than the polar modality due to the reason that the off-the-shelf VGGFace [145] feature extractor is pretrained on the visible face dataset.

6.3 Summary

We proposed a novel self-attention guided network for synthesizing thermal and visible faces for the task of cross-spectral face matching. Given visible probe images, we synthesize the corresponding thermal images. Similarly, given thermal probe images, we synthesize the visible images. Features are then extracted from the original and the synthesized images. Their fused feature representations are then used for verification. The generators are based on the self-attention guided networks. Various experiments on the ARL polarimetric thermal dataset were conducted to show the significance of the proposed approach. Furthermore, an ablation study was conducted to show the improvements achieved by the proposed fusion approach.

Chapter 7

Heterogeneous Face Frontalization via Domain Agnostic Learning

As discussed earlier, face recognition is a challenging problem which has been actively researched over the past few decades. However, existing methods are specifically designed for recognizing face images that are collected in the visible spectrum. In applications such as night-time surveillance, we are faced with the scenario of identifying a face image acquired in thermal domain by comparing it with a gallery of face images that are acquired in the visible domain. Existing DCNN-based visible face recognition methods will not perform well when directly applied to the problem of thermal to visible face recognition due to the significant distributional shift between the thermal and visible domains. In order to bridge this gap, various cross-domain face recognition algorithms have been proposed [1, 26–28, 35, 60–62, 66, 68, 171]. In particular, synthesis-based methods have gained

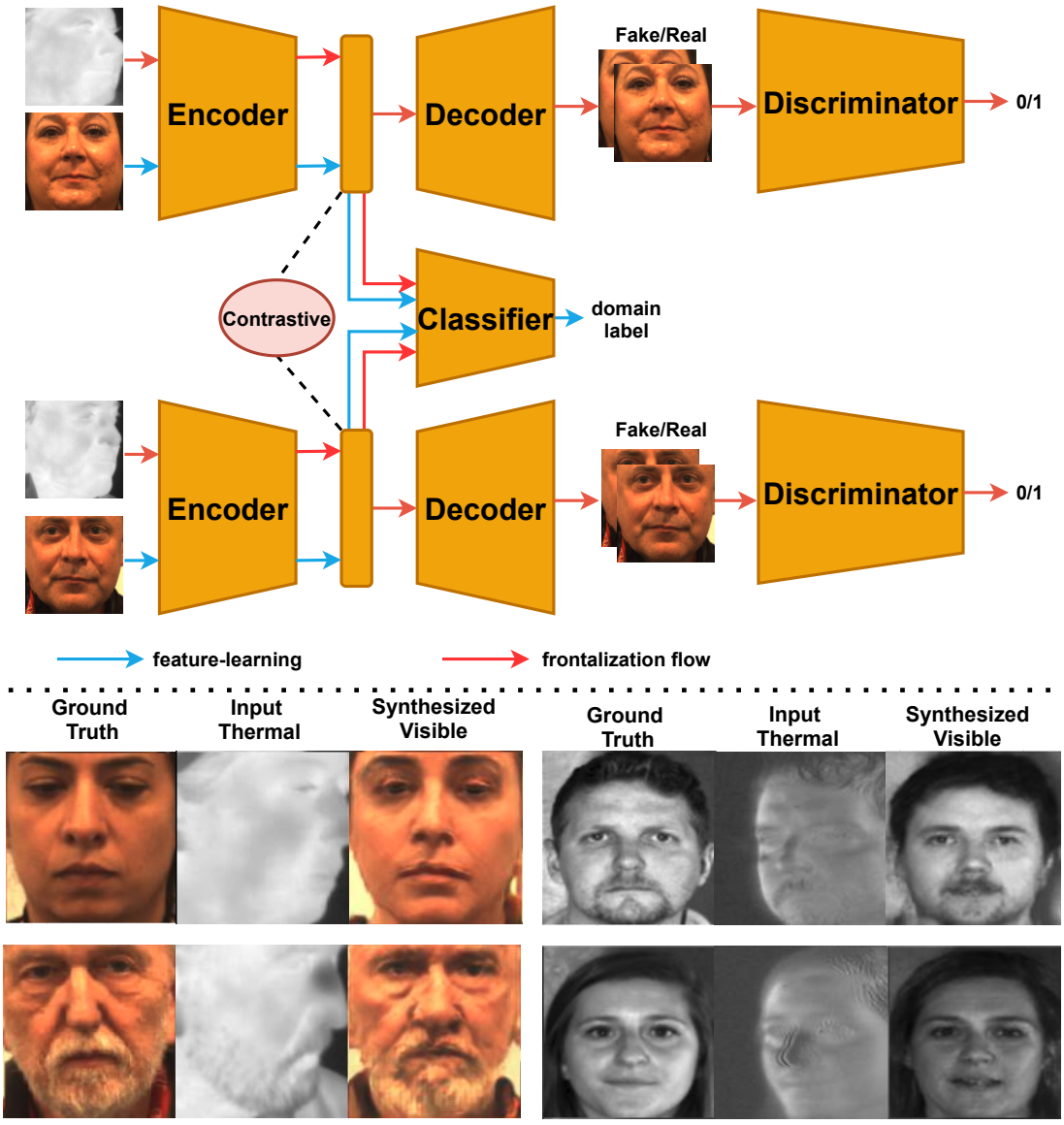


Figure 7.1: An overview of the proposed heterogeneous face frontalization method. Frontalization flow aims to reconstruct a visible frontal face from a thermal profile face. Feature-learning flow aims to learn domain-agnostic features that imitate the domain discrepancy between input and output images. To enhance feature discrimination, the latent features are regularized by a contrastive constraint during training.

a lot of traction in recent years [35]. Given a thermal face image, the idea is to synthesize the corresponding face image in the visible domain. Once the visible image is synthesized, any DCNN-based visible face recognition network can be leveraged for identification.

One of the main limitations of the existing synthesis-based models for cross-modal face recognition is that they do not perform well on thermal faces with large pose variations [35]. Face frontalization is an extensively studied problem in the computer vision and biometrics communities. Various methods have been developed for frontalization [13–15,72,73,75,76, 79–84]. However, existing face frontalization methods' performance degrades significantly on thermal faces since they are specifically designed for frontalizing visible face images. In order to deal with this problem, heterogeneous face frontalization methods are needed in which a model takes a thermal profile face image and generates a frontal visible face. This is an extremely difficult problem due to the large domain as well as large pose discrepancies between the two modalities. Despite its applications in biometrics and surveillance, this problem is relatively unexplored in the literature.

We propose a domain agnostic learning-based generative adversarial network (DAL-GAN) [9] with dual-path training architecture which can synthesize frontal views in the visible domain from thermal faces with pose variations. Figure 7.2 gives a simplified overview of the proposed heterogeneous frontalization framework. The model is trained on two flows: frontalization flow and feature-learning flow. The frontalization flow aims to synthesize a visible frontal face image from a thermal profile face image. The feature learning flow aims to learn domain-agnostic features that help in reconstructing better vis-

ible faces. DAL-GAN consists of a generator with an auxiliary classifier and two discriminators which capture both local and global texture discriminations for better synthesis. A contrastive constraint is enforced in the latent space of the generator with the help of a dual-path training strategy, which improves the feature vector's discrimination. Finally, a multi-purpose loss function is utilized to guide the network in synthesizing identity-preserving cross-domain frontalization. We conduct extensive experiments to demonstrate that DAL-GAN can generate better quality frontal views compared to the other baseline methods. Figure 7.2 shows sample outputs from the proposed network.

7.1 Proposed Method

Given a thermal face image $\mathbf{x}$ with a significant pose variation, our objective is to synthesize the corresponding frontal face image $\hat{\mathbf{y}}$ in the visible domain. The generated frontal face image should be photo-realistic and identity-preserving. In order to address this cross-spectrum face frontalization problem, we propose a dual-path network architecture which learns domain-agnostic features via a contrastive learning strategy. In what follows, we present the proposed network in detail.

7.1.1 Networks Architecture

Figure 7.2 gives an overview of the proposed dual-path architecture. Each path contains a generator F and two discriminators D^g, D^l. The two generators share weights. Another

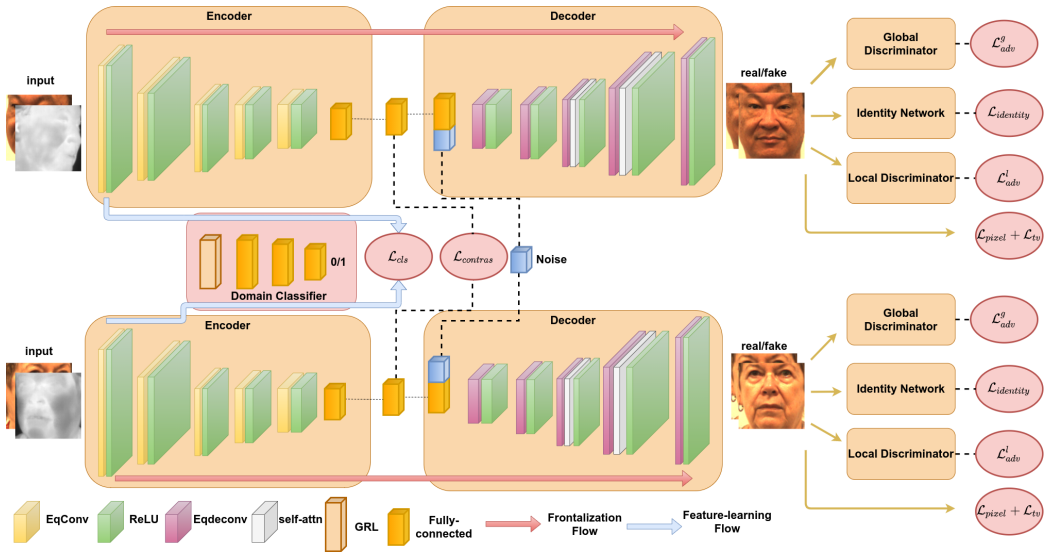


Figure 7.2: Illustration of the proposed dual-path architecture. Two weight-shared identical generators are employed in each path. Both the face **frontalization flow** and domain-agnostic **feature-learning flow** are implemented during training. For frontalization, a proper combination of multiple losses are utilized which contain the multi-scale pixel loss $\mathcal{L}_{pixel}$, identity loss $\mathcal{L}_{id}$, global and local adversarial losses $\mathcal{L}^{g}_{adv}$ and $\mathcal{L}^{l}_{adv}$, total variation loss $\mathcal{L}_{tv}$ as well as contrastive loss $\mathcal{L}_{contras}$. Additionally, another domain classifier network is utilized for learning domain-agnostic feature which is optimized by the classification loss $\mathcal{L}_{cls}$ with a gradient reversal layer (GRL).

domain classifier C is utilized for learning domain-agnostic features. More details about the network architecture can be found in the supplementary file.

7.1.1.1 Generator

Each generator F consists of an encoder-decoder structure where the encoder E aims to extract domain-agnostic features from the input thermal face while the decoder G aims to reconstruct a frontal face image in the visible domain. In this work, we adopt the generator architecture from DA-GAN [13] with the following modifications.

We remove all batch normalization layers in DA-GAN because they were originally introduced to eliminate the covariate shift. However, recent studies have shown that covariate shift often does not exist in GANs [38,133,134]. Instead we use feature vector equalization to prevent the escalation of feature magnitudes. Feature equalization also helps the network converge smoothly during training. Given the original feature vector $a_{m,n}$ at pixel location (m, n), the equalization is defined as follows

$$\mathbf{b}_{m,n} = \mathbf{a}_{m,n} / \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} \mathbf{a}_{m,n}^{i} + \epsilon}, \tag{7.1}$$

where $\mathbf{b}_{m,n}$ is the normalized vector and N is the total number of features. We set ϵ equal to 10^{-8}. In our experiments, we find that feature equalization also allows us to use a higher initial learning rate which in turn helps in accelerating training.

7.1.1.2 Discriminator

A single discriminator may not be able to capture both global and local facial textures [15, 82]. Therefore, we employ two identical discriminators (D_l, D_g) to learn global and local discrimination respectively, as shown in Figure 7.3. In particular, the local components of frontal face $\hat{\mathbf{y}}_f$ are extracted by a predefined off-the-shelf model F_m [172]. In this work, the key components (eyes, nose, lips, brow) are extracted to learn the local facial structure and the entire face image is used for learning the global texture. Mathematically, we define a mask M to extract the key components of a ground-truth visible image $\mathbf{y}$ as:

$$\mathbf{M} = F_m(\mathbf{y}).$$

With the help of this mask, we can obtain local regions of a real/fake image by the element-wise product $\odot$ between the mask and the real/fake image. The local/global discriminators are optimized by the following loss function $\mathcal{L}^l_{adv}$/$\mathcal{L}^g_{adv}$ separately

$$\begin{aligned}
\mathcal{L}^g_{adv} &= \mathbb{E}[D_g(\mathbf{y})] - \mathbb{E}[(D_g(\hat{\mathbf{y}}))] - \\
&\lambda_{gp}\mathbb{E}[(\|\nabla_{\mathbf{y}^\star} D_g(\mathbf{y}^\star)\|_2 - 1)^2], \\
\mathcal{L}^l_{adv} &= \mathbb{E}[D_l(\mathbf{M}\odot\mathbf{y})] - \mathbb{E}[(D_l(\mathbf{M}\odot\hat{\mathbf{y}}))] - \\
&\lambda_{gp}\mathbb{E}[(\|\nabla_{\mathbf{M}\odot\mathbf{y}^\star} D_l(\mathbf{M}\odot\mathbf{y}^\star)\|_2 - 1)^2], \\
\mathcal{L}_{adv} &= \mathcal{L}^g_{adv} + \lambda_l \cdot \mathcal{L}^l_{adv}.
\end{aligned} \tag{7.2}$$

Here, $\mathbf{y}^\star_f$ is sampled uniformly along a straight line between a pair of real image $\mathbf{y}_f$ and the generated image $\hat{\mathbf{y}}_f$ [173]. The discriminators attempts to maximize those objective, while the generator attempts to minimize it. We set $\lambda_{gp} = 10, \lambda_l = 0.1$ in our experiments.

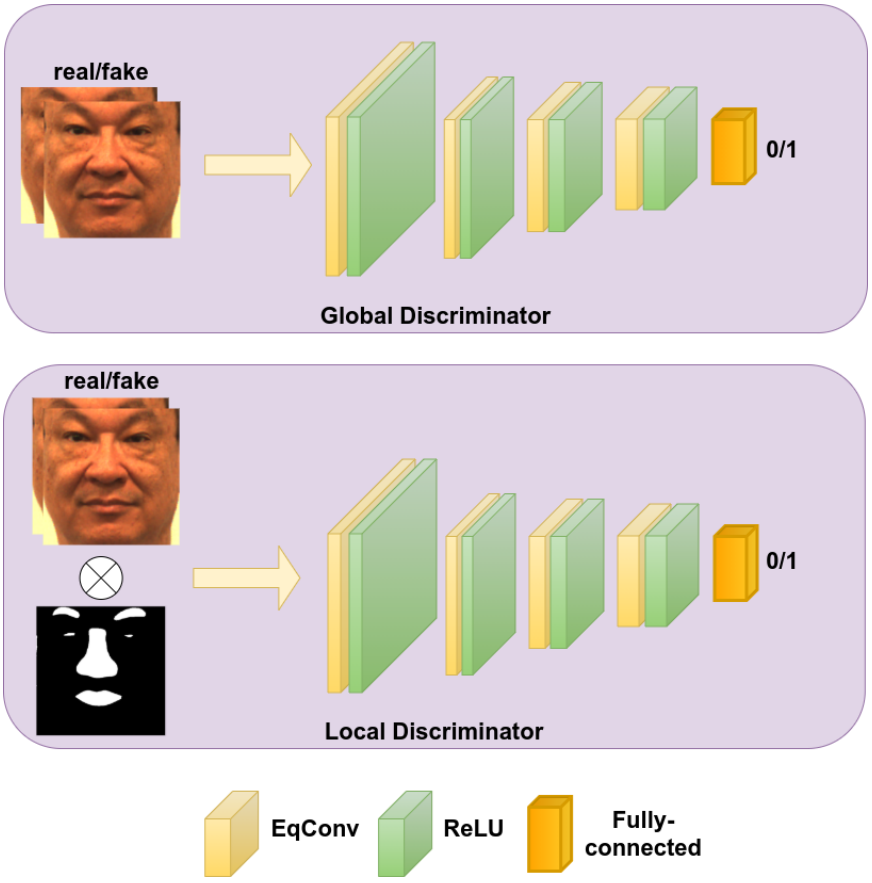


Figure 7.3: Illustration of the global and local discriminators.

7.1.2 Objective Function

The loss function we use to train the proposed network consists of a combination of contrastive loss, gradient reversal layer-based domain adaptive loss and frontalization loss.

Frontalization Loss: To capture texture information between the real and fake images at multiple scalese, we utilize a multi-scale pixel loss as follows

$$\mathcal{L}_{pixel} = \sum_{s=1}^{3} \|\hat{\mathbf{y}}^s - \mathbf{y}^s\|_1, \tag{7.3}$$

where s corresponds to the resolution scale. In our work, we utilize three different resolution scales: 32×32, 64×64 and 128×128.

Identity preserving synthesis is important for heterogeneous face recognition. To achieve this, we utilize the identity loss as follows

$$\mathcal{L}_{id} = \|F_v(\hat{\mathbf{y}}) - F_v(\mathbf{y})\|_2, \tag{7.4}$$

where F_v is the pre-trained VGGFace [145] model which is used to extract features. By minimizing the feature distance between the real and fake images, the generator is optimized to synthesize identity-preserving images.

In order to reduce the artifacts in the synthetic images, a total variation regularization $\mathcal{L}_{tv}$ is also applied as that in [160]. Hence, the total loss for the frontalization flow is defined as follows

$$\mathcal{L}_{front} = \mathcal{L}_{pixel} + \lambda_I \cdot \mathcal{L}_{id} + \lambda_{adv} \cdot \mathcal{L}_{adv} + \lambda_{tv} \cdot \mathcal{L}_{tv}, \tag{7.5}$$

where $\lambda_I, \lambda_{adv}, \lambda_{tv}$ are parameters.

Domain Adaptive Loss: To learn domain-agnostic features, one domain classifier network C with a gradient reversal layer [174] is employed in our network. The domain classifier network is a simple 3-layer multi-linear perceptron (MLP) neural network with the same equalization as in Eq (7.1). Given both thermal and visible face images, the classifier aims to correctly estimate which spectrum the input image belongs to. When back-propagating, the gradient reversal layer flips the gradients which leads to domain-agnostic features in the encoder. The network parameters θ_E corresponding to the encoder network are updated as follows

$$\begin{aligned} \mathcal{L}_{cls} = \frac{1}{N}\sum_{i=0}^{N-1} -\mathbf{k}_i \log(C(E(\mathbf{x}_i))) - (1-\mathbf{k}_i)\log(1-C(E(\mathbf{x}_i))) \\ \theta'_E = \theta_E - \mu\left(\frac{\partial \mathcal{L}_{front}}{\partial \theta_E} - \lambda_{grl}\frac{\partial \mathcal{L}_{cls}}{\partial \theta_E}\right), \end{aligned} \tag{7.6}$$

where $\mathbf{x}_i$ is the i-th input sample, μ is the learning rate, θ'_E and θ_E are the updated and original parameters, respectively. $\mathbf{k}_i$ is the domain index $[0, 1]$. λ_{grl} is set equal to 0.01 in our experiments.

Contrastive Loss: Finally, the contrasitive loss [175] is used to enhance the discriminability of the latent features. It is defined as follows

$$\begin{aligned} \mathcal{L}_{Contras} = \mathbf{l} \cdot \|E(\mathbf{x}^1) - E(\mathbf{x}^2)\|_2 + \\ (1-\mathbf{l}) \cdot \max(0, m - \|E(\mathbf{x}^1) - E(\mathbf{x}^2)\|_2), \end{aligned} \tag{7.7}$$

where $\mathbf{x}^1, \mathbf{x}^2$ are two profile face images input to the dual-path architecture. The label vector $\mathbf{l}$ indicates whether they contain the same ($\mathbf{l} = 1$) or different ($\mathbf{l} = 0$) identity. The hyper-parameter $m = 1.2$ indicates the margin.

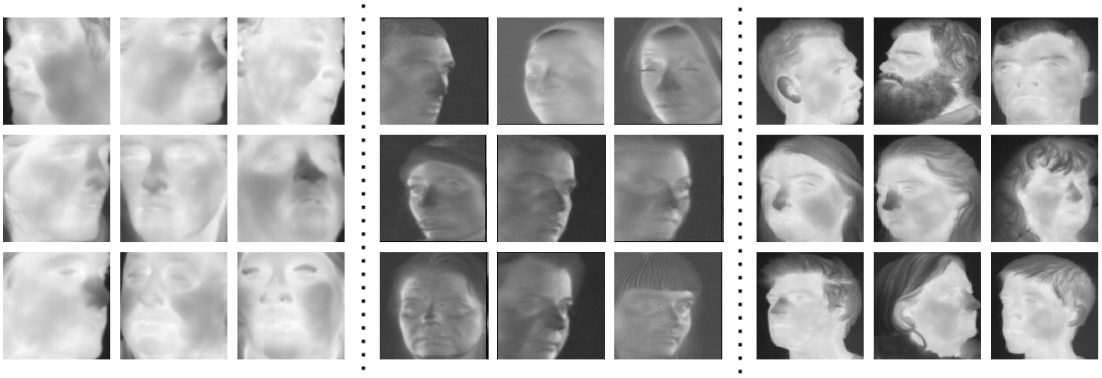
Total Objective: In order to reduce the artifacts in the synthetic images, a total variation regularization $\mathcal{L}_{tv}$ is also applied as that in [160]. The overall loss function used to train the proposed heterogeneous frontalization network is as follows

$$\mathcal{L} = \mathcal{L}_{front} + \lambda_C \cdot \mathcal{L}_{contras} + \lambda_{cls} \cdot \mathcal{L}_{cls}, \tag{7.8}$$

where $\lambda_{con}, \lambda_{cls}$ are regularization parameters.

7.2 Experiments

Datasets: We evaluate the proposed heterogeneous face frontalization network on three publicly available datasets: DEVCOM Army Research Laboratory Visible-Thermal Face Dataset (ARL-VTF) [176], ARL Multimodal Face Database (ARL-MMFD) [34, 35, 140] and TUFTS Face [177]. Sample images from these datasets are shown in Figure 7.4.



ARL-VTF [176] ARL-MMFD [35] TUFTS [177]

Figure 7.4: Input profile thermal images sampled from three datasets respectively.

ARL-VTF Dataset: The ARL-VTF dataset [176] consists of 500,000 images from 395 subjects (295 for training, 100 for testing). This dataset contains a large collection of paired visible and thermal faces. Variations in baseline, expressions, pose, and eye-wear are included. We first crop the images based on the given ground-truth bounding-box annotations for conducting experiments. The proposed model is trained on this dataset by the predefined development and testing splits [176]. The final experimental results are based on the average of the predefined 5 splits.

ARL-MMFD Dataset: Additionally, we evaluate the proposed model on Volume III of the ARL-MMFD dataset [35]. This dataset was collected by ARL cross 11 sessions over 6 days. It contains 5419 paired polarimetric thermal and visible images from 121 subjects with significant variations in expression, illuminations, pose, glasses, etc. We select the polarimetric thermal profile and visible frontal face image pairs corresponding to both neutral and expressive faces for conducting experiments on this dataset. In particular, we randomly select images from 90 identities for training and the images from the remaining identities for testing. So, there are no overlapping images between training and testing sets. We use the original aligned visible and polarimetric thermal images for training and testing without any preprocessing. The final experimental results are based on the average of five random splits.

TUFTS Face Dataset: Finally, the proposed model is also evaluated on a recently pub-

lished TUFTS Face Dataset [177]. TUFTS is a multi-modal dataset which contains more than 10,000 images from 113 individuals from more than 15 countries. For conducting experiments, we only select thermal and visible images from this multi-modal dataset. Hence, there are more than 1000 images from over 100 subjects with different pose and expression variations. This dataset is very challenging due to a large number of pose and expression variations and only a few images per variation are available in the dataset. Images from randomly selected 89 individuals are used for training and images from the remaining 23 subjects are used for testing. The raw images are used to train and test without any preprocessing. In particular, profile images in the thermal domain and the frontal images (both neural and expression) in the visible domain are utilized for training the models.

Implementation Details: All the images in this work are resized to 128×128 and the image intensity is scaled into [0, 1]. The features from VGGFace [178] average pooling layer are extracted from the synthesized visible image and the similarity for verification is calculated based on the cosine distance. In all the experiments, the hyper-parameters are set as $\lambda_I = 10, \lambda_{adv} = 1, \lambda_C = 0.01, \lambda_{tv} = 1e^{-4}$. The learning rate is initially set as 0.01 and the batch size is 8. We train our model 10, 100, and 400 epochs on ARL-VTF, ARL-MMFD and TUFTS datasets, respectively.

We compare the proposed method with the following state-of-the-art facial frontalization methods: TP-GAN [15]; PIM [82]; M2FPA [86]; DA-GAN [13]. Besides, we add pix2pix [159] as a baseline, which is an image-to-image translation method between two

domains. We followed the same network settings as mentioned in the original papers and tried our best to finetune the training parameters. Note that TP-GAN [15] and PIM [82] require landmarks on input profile thermal face images. We manually label the landmarks on thermal faces in the TUFTS Face Dataset [177] and use officially provided landmarks in the ARL-VTF dataset [176]. We estimate the landmarks by MTCNN [109] on the images in the ARL-MMFD dataset [35].

7.2.1 Experimental Results

ARL Visible-Thermal Face Dataset: The evaluation is based on the following two protocols: (1) Gallery G_VB0- to Probe P_PTP0. (2) Gallery G_VB0- to Probe P_PTP-. The images in Gallery G_VB0- are the facial baseline images in the visible domain without eye-glasses occlusion. The images in Probe P_PTP0 are the facial profile images in the thermal domain without eye-glasses occlusion. The images in Probe P_PTP- are the facial profile images in the thermal domain with eye-glasses token-off.

We evaluate our model with the other baseline methods both qualitatively and quantitatively. Visual frontalization results are shown in Figure 7.5. As can be seen from this figure, the proposed method is able to synthesize more photo-realistic and identity-preserving images compared with the other methods. Other methods fail to synthesize high-quality and identity-preserving images. Additionally, Table 7.1 quantitatively compares the verification performance of different methods. Our method achieves around 2% and 5% improvements on both the AUC and EER scores in two protocols, respectively when compared with the

Table 7.1: Verification performance comparison on the ARL-VTF dataset [176].

	Probe	00_pose				10_pose			
Gallery	Metric	AUC	EER	FAR=1%	FAR=5%	AUC	EER	FAR=1%	FAR=5%
Gallery 0010	Raw	54.88	46.38	2.16	8.33	56.10	45.98	1.06	8.53
	Pix2Pix [159]	5.04	46.84	2.17	8.90	57.95	44.84	1.79	14.47
	TP-GAN [15]	64.21	40.12	3.31	12.11	67.41	36.78	3.89	13.84
	PIM [82]	68.69	36.58	5.43	16.56	73.31	32.70	5.96	20.42
	M2FPA [86]	74.99	32.33	5.73	20.26	76.99	29.84	9.51	23.35
	DA-GAN [13]	75.58	31.18	6.85	22.23	75.76	30.69	8.40	23.62
	Ours	77.48	29.08	8.20	25.89	82.18	25.11	10.82	30.64

Table 7.2: Verification performance comparison corresponding to the ablation study.

	Probe	00_pose				10_pose			
Gallery	Metric	AUC	EER	FAR=1%	FAR=5%	AUC	EER	FAR=1%	FAR=5%
Gallery 0010	Baseline	56.21	45.99	2.07	9.04	59.20	43.13	2.73	9.71
	w/ Multi-scale $\mathcal{L}_1$	58.07	44.82	2.13	10.27	62.01	40.91	3.84	11.06
	w/ $\mathcal{L}_{id}$	70.26	34.26	4.90	16.26	76.56	29.67	6.70	23.84
	w/ self-attn	71.99	34.05	5.22	19.72	76.60	29.94	6.62	24.00
	w/ $\mathcal{L}^l_{adv}$	72.58	33.54	6.21	20.07	76.58	28.44	6.66	24.18
	w/ Eq.(7.1)	77.13	29.16	6.41	21.35	80.05	26.26	7.11	29.40
	w/ $\mathcal{L}_{cls}$	77.18	29.07	7.43	22.20	80.56	26.46	9.24	29.90
	w/ $\mathcal{L}_{contras}$ (ours)	77.48	29.08	8.20	25.89	82.18	25.11	10.82	30.64

recent DA-GAN [13]. Additionally, the proposed model also achieves $2\% \sim 5\%$ improvements on the True Accept Rate (TAR) at False Accept Rates (FAR) 1% and 5%.

ARL Multimodal Face Database: We evaluate the proposed model on Volume III of ARL-MMFD [35]. Qualitative and quantitative results corresponding to this dataset are shown in Figure 7.6 and Table 7.3, respectively. As can be seen from Table 7.3, the proposed model surpasses the best performing baseline model by 2.4% and 1.5% in AUC and EER scores, respectively. Furthermore, the proposed model surpasses 7.7% and 1.3% when compared with the DA-GAN [81] at FAR=1% and FAR=5%, respectively.

From Figure 7.6, we can see that our model is able to synthesize more photo-realistic images while preserving the identity better than the other frontalization models. For each

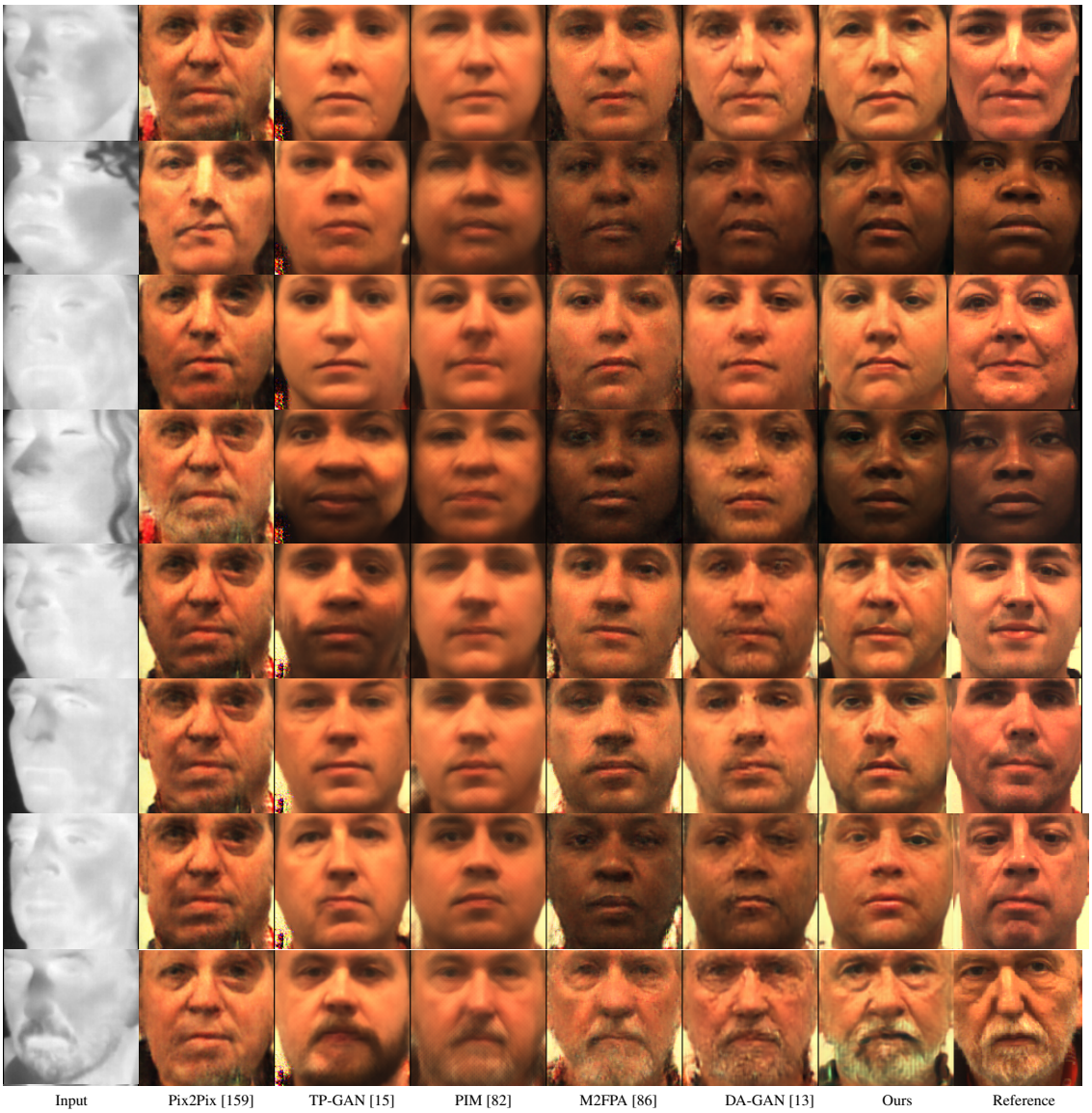


Figure 7.5: Cross–domain face frontalization comparison on the ARL-VTF [176] dataset.

Table 7.3: Verification performance comparison on the ARL-MMFD dataset [35].

Method	AUC	EER	FAR=1%	FAR=5%
Raw	64.12	40.51	4.21	17.74
Pix2Pix [159]	73.60	31.96	11.16	26.45
TP-GAN [15]	76.15	30.89	6.26	19.03
PIM [82]	80.89	26.86	9.83	27.11
M2FPA [86]	85.58	22.27	13.32	37.58
DA-GAN [13]	86.26	21.56	14.74	33.17
Ours	88.61	20.21	16.07	40.93

subject in this dataset, the pose variations mainly cover from $-60^{\circ} \sim 60^{\circ}$, while the illumination variation is various. Hence, the method like PIM [82] fails to synthesize photo-realistic images due to the illumination artifacts. In addition, our proposed model is trained based on contrastive loss which helps the network provide better results even on this smaller dataset.

TUFTS Face Database: Finally, the proposed model is also evaluated on the TUFTS Face dataset [177]. The performances are reported in quantitative as shown in Table 7.4. We can observe our proposed method surpasses the previous baselines and achieves around 3% improvement in AUC and EER scores.

Figure 7.6: Cross-domain face frontalization comparison on the ARL-MMFD [35] dataset.

Table 7.4: Verification performance comparison on the TUFTS Face Database [177].

Method	AUC	EER	FAR=1%	FAR=5%
Raw	67.55	37.88	5.11	16.11
Pix2Pix [159]	69.71	35.31	5.44	21.66
TP-GAN [15]	70.93	35.32	6.46	18.77
PIM [82]	72.84	34.10	8.77	21.00
M2FPA [86]	75.07	31.22	8.33	23.44
DA-GAN [13]	75.24	31.14	10.44	26.22
Ours	78.68	28.38	10.44	27.11

7.2.2 Ablation Study

In this chapter, we analyze how each part of the proposed model contributes to the final performance. We choose the global UNet [15] generator and the discriminator D_g as the baseline model. In particular, we analyze the contribution of each loss function and the equalization in Eq (7.1). We conduct these ablation experiments on the ARL-VTF dataset and show the results both qualitatively and quantitatively in Figure 7.7 and Table 7.2, respectively.

As shown in Table 7.2, the quantitative verification results are improved by consecutively adding different components. Figure 7.7 shows the corresponding synthesized samples. Based on this ablation study, we can see that $\mathcal{L}_{id}$ significantly preserves the identity in the synthesized image. Feature equalization Eq. (7.1) significantly improves the image quality. Finally, $\mathcal{L}_{cls}$ and $\mathcal{L}_{contras}$ improve the verification performance.

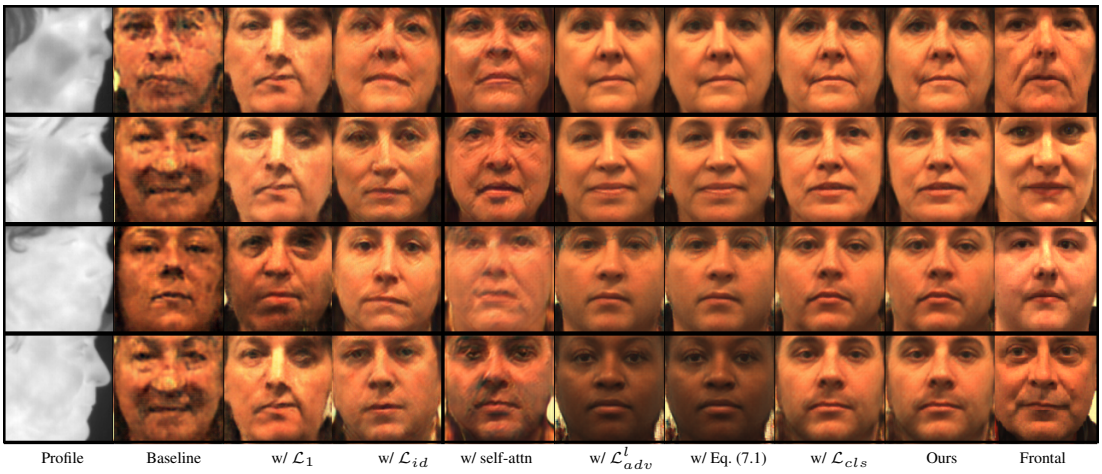


Figure 7.7: Results corresponding to the ablation study.

7.2.3 Pose Invariant Representation

In Figure 7.9, we show the synthesized frontal faces from the proposed model with a range of yaw poses on the ARL-MMFD dataset. Yaw poses in this dataset mainly cover $0^\circ \sim 45^\circ$. Given arbitrary yaw pose polarimetric thermal images, we can observe that the synthesized visible frontal images maintain a good pose-invariant representation. Additionally, we also conduct this analysis on the ARL-VTF dataset as shown in Figure 7.9. From these two figures, we can see that the proposed model learns pose-invariant frontal representation in the visible domain when given arbitrary pose thermal input images.

7.3 Summary

In this work, we proposed a novel heterogeneous face frontalization model which generates frontal visible faces from profile thermal faces. The generator contains a gradient

Figure 7.8: Synthesized frontal images corresponding to a range of yaw poses on the ARL-VTF dataset.

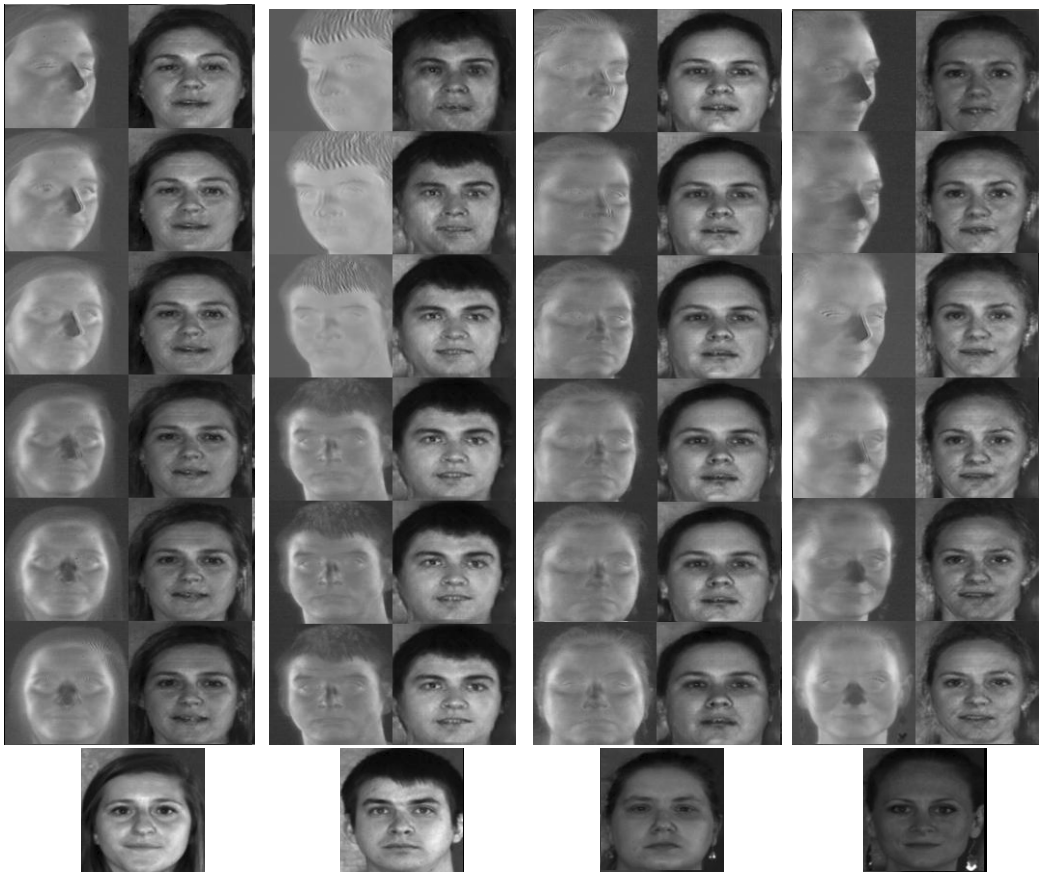


Figure 7.9: Synthesized frontal images corresponding to a range of yaw poses on the ARL-MMFD dataset.

reversal layer-based classifier for domain-agnostic feature learning and a pair of local and global discriminators for better synthesis. Additionally, another contrastive constraint is enforced by a dual-path training strategy for learning discriminative latent features. Quantitative and visual experiments conducted on three real thermal-visible datasets demonstrate the superiority of the proposed method when compared to other existing methods. Additionally, an ablation study was conducted to demonstrate the improvements obtained from different modules and loss functions.

Chapter 8

GP-GAN: Gender Preserving GAN for Synthesizing Faces from Landmarks

Facial landmarks can be regarded as the most compressed representation of a face due to the fact that very few number of points are required to capture the landmark locations. In spite of the incredibly low number of keypoints, they are known to preserve important information about the face such as pose, gender [179] and structure [180–182]. Success of facial analysis tasks using just landmark keypoints is essential from the perspective of memory management and information privacy. Considering that size of landmarks is an order of magnitude smaller as compared to the image size, it will result in significant savings in terms of memory. Essentially, we are now able to store only landmark key points and throw away face image for a particular application. In addition, landmark information can be safely stored, transported, and distributed without potential violation of human privacy

and confidentiality. Motivated by these reasons, it would be interesting to understand how landmarks can be exploited for performing high-level facial analysis tasks in the absence of corresponding face images.

Several researchers have demonstrated that facial landmarks can be used in many face analysis tasks such as face recognition [149, 183, 184], facial attribute inference [185], age estimation [180], gender recognition [179] and expression analysis [186]. However, these methods operate on a small set of keypoints due to which their performance is severely limited. To overcome this problem, we propose a novel solution that involves synthesis of faces from landmark points using the recently popular generative models [119, 166, 187–191]. While, several methods [185, 192–194] have been proposed in the literature for landmark detection, the inverse problem of synthesizing faces from their corresponding landmarks is a largely unexplored problem. We believe that using synthesized faces will result in better recognition performances as they leverage the capabilities of generative models to accentuate information present in landmarks. Apart from their use in high-level facial analysis tasks, these generative methods can be used to create virtually unlimited stochastic samples by conditioning on both landmarks and a stochastic noise vector enabling us to augment existing datasets for large scale learning [195].

In this work, generative models are exploited to synthesize faces from landmarks in an attempt to accentuate information (gender in particular) present in the landmarks. Cao *et al.* [179] specifically address the question if facial metrology can be used to predict gender and they further go on to demonstrate that gender recognition using landmarks achieves

reasonable performance. This is remarkable considering the fact that only 68 keypoints are used to predict gender of the face represented by these keypoints. However, generating faces from landmarks will enable us to achieve further improvement in performance as this process will leverage generative models to learn the distribution of landmarks and their mappings to the respective faces. While recognition of other attributes like ethnicity, pose, identity, etc. can all be improved, in this work, we specifically focus on the gender attribute. To this end, we propose Gender Preserving Generative Adversarial Network (GP-GAN) to generate faces from their respective landmarks (as shown in Fig. 8.1). To further enhance the network's performance, it is guided by perceptual loss and a gender preserving loss in addition to adversarial loss.

8.1 Proposed Method

Given an application where only facial landmarks are available, we explore how to leverage information preserved in these keypoints. To this end, we propose to model the joint distribution of facial landmarks and corresponding face images [1] using generative modeling. Inspired by the success of GANs [187], we explore adversarial networks in this work for synthesizing faces from landmark keypoints. GANs, motivated by game theory, consist of two competing networks: generator G and discriminator D. The goal of GAN is to train G to produce samples from training distribution such that the synthesized samples are indistinguishable from actual distribution by discriminator D. Conditional GAN is

[1] Face images are available only during training

another variant where the generator is conditioned on additional variables such as discrete labels [55], text [19] and images [54]. The objective function of a conditional GAN is defined as follows

$$L_{cGAN}(G,D) = E_{x,y\sim Pdata(x,y)}[\log D(x,y)]+ \\ E_{x\sim Pdata(x),z\sim p_z(z)}[\log(1-D(x,G(x,z)))], \tag{8.1}$$

where y, the output image, and x, the observed image, are sampled from distribution $Pdata(x,y)$ and they are distinguished by the discriminator, D. While for the generated fake $G(x,z)$ sampled from distributions $x \sim Pdata(x), z \sim p_z(z)$ would like to fool D.

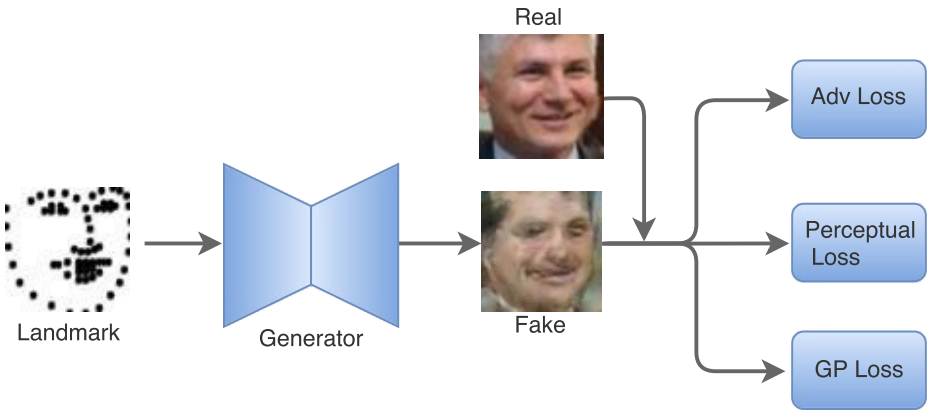


Figure 8.1: Overview of the proposed GP-GAN method for synthesizing faces from landmarks. In addition to adversarial loss function, the generator sub-network is guided by a perceptual loss and a gender preserving loss.

As shown in Fig. 5.1, the proposed network consists of a generator sub-network G (based on U-net [29] and DenseNet [196] architecture) conditioned on a facial landmark image and a patch-based discriminator sub-network D. G takes landmark as input and attempts to generate corresponding face image, while D attempts to distinguish between real

and synthesized images. The two sub-networks are trained iteratively. In addition to the adversarial loss, we propose to guide the generator using three other loss functions: perceptual loss based on VGG-16 architecture [161], gender preserving loss and L_1 reconstruction error.

8.1.1 Generator

Deeper networks are known to better capture high-level concepts, however, the vanishing gradient problem affects convergence rate as well as the quality of convergence. Several works have been developed to overcome this issue among which U-Net [29] and DenseNet [196] are of particular interest. While U-Net incorporates longer skip connections to preserve low-level features, DenseNet employs short range connections within micro-blocks resulting in maximum information flow between layers in addition to an efficient network. Motivated by these two methods, we propose UDeNet for the generator sub-network G in which, the U-Net architecture is seamlessly integrated into the DenseNet network in order to leverage advantages of both the methods. This novel combination enables more efficient learning and improved convergence quality.

A set of 3 dense-blocks (along with transition blocks) are stacked in the front, followed by a set of 5 dense-block layers (transition blocks). The initial set of dense-blocks are composed of 6 bottleneck layers. For efficient training and better convergence, symmetric skip connections are involved into the generator sub-network, similar to [197]. Details regarding the number of channels for each convolutional layer are as follows:

C(64)-M(64)-D(256)-T(128)-D(512)-T(256)-D(1024)-T(512)-D(1024)-DT(256)-D(512)-DT(128)-D(256)-DT(64)-D(64)-D(32)-D(32)-DT(16)-C(3),

where C(K) is a set of K-channel convolutional layers followed by batch normalization and ReLU activation. M is max-pooling layer. D(K) is the dense-block layer with K-channel output, T(K) is transition layer with K-channel output for downsampling. DT(K) is similar to T(K) except for transposed convolutional layer instead of convolutional layer for upsampling.

8.1.2 Discriminator

Motivated by [54], patch-based discriminator D is used and it is trained iteratively along with G. The primary goal of D is to learn to discriminate between real and synthesized samples. This information is backpropagated into G so that it generates samples that are as realistic as possible. Additionally, patch-based discriminator ensures preserving of high-frequency details which are usually lost when only L1 loss is used. All the convolutional layers in D have a filter size of 4×4. Details regarding the number of channels for each convolutional layer are specified in Fig. 5.1.

8.1.3 Objective function

The network parameters are learned by minimizing the following objective function:

$$L = L_A + \lambda_P L_P + \lambda_C L_C + \lambda_1 L_1, \tag{8.2}$$

where L_A is the adversarial loss, L_P is the perceptual loss, L_C is the gender preserving loss and L_1 is the loss based on L_1-norm between the target and reconstructed image, λ_P, λ_C and λ_1 are weights respectively for perceptual loss, gender preserving loss and L_1 loss.

Adversarial loss: Adversarial loss is based primarily on the discriminator sub-network D. Given a set of N synthesized faces, $\{\hat{x}_i\}_{i=1}^{N}$, the entropy loss from D that is used to learn the parameters of G is defined as:

$$L_A = -\frac{1}{N}\sum_{i=1}^{N} log(D(\hat{x}_i)), \tag{8.3}$$

Perceptual loss: Johnson *et al.* [160] introduced the perceptual loss function for style transfer and super-resolution. Instead of relying only on L_1 or L_2 reconstruction error, they learn the network parameters using errors between high-level image feature representations extracted from a pre-trained convolutional neural network. Similar to their work, pre-

trained VGG-16 [161] network is used to extract high-level features (conv4_3 layers) and the L_1 distance between these features of real and fake images is used to guide the generator G. The perceptual loss function is defined as:

$$L_P = ||V(\hat{x}) - V(x)||_1, \tag{8.4}$$

where, x and $\hat{x}$ indicate real and fake images, respectively and V is a particular layer of the VGG-16 network.

Gender preserving loss: Inspired largely by the perceptual loss, we define a gender preserving loss. As indicated by the name, this function measures the error in terms of gender attribute of the synthesized image as compared to that of real image. It is defined as:

$$L_C = -\frac{1}{N}\sum_i (C(x_i)\log(C(\hat{x}_i)) + (1 - C(x_i))\log(C(\hat{x}_i)),$$

where C represents a pre-trained gender classification network. In this work, C is constructed using the standard VGG-16 network in which, the convolutional layers are retained and the fully connected layers are replaced by a new set of layers as shown in Fig. 5.1. This network is trained by minimizing the standard binary cross entropy error.

L1 loss: L1 loss measures the reconstruction error between the synthesized face image and

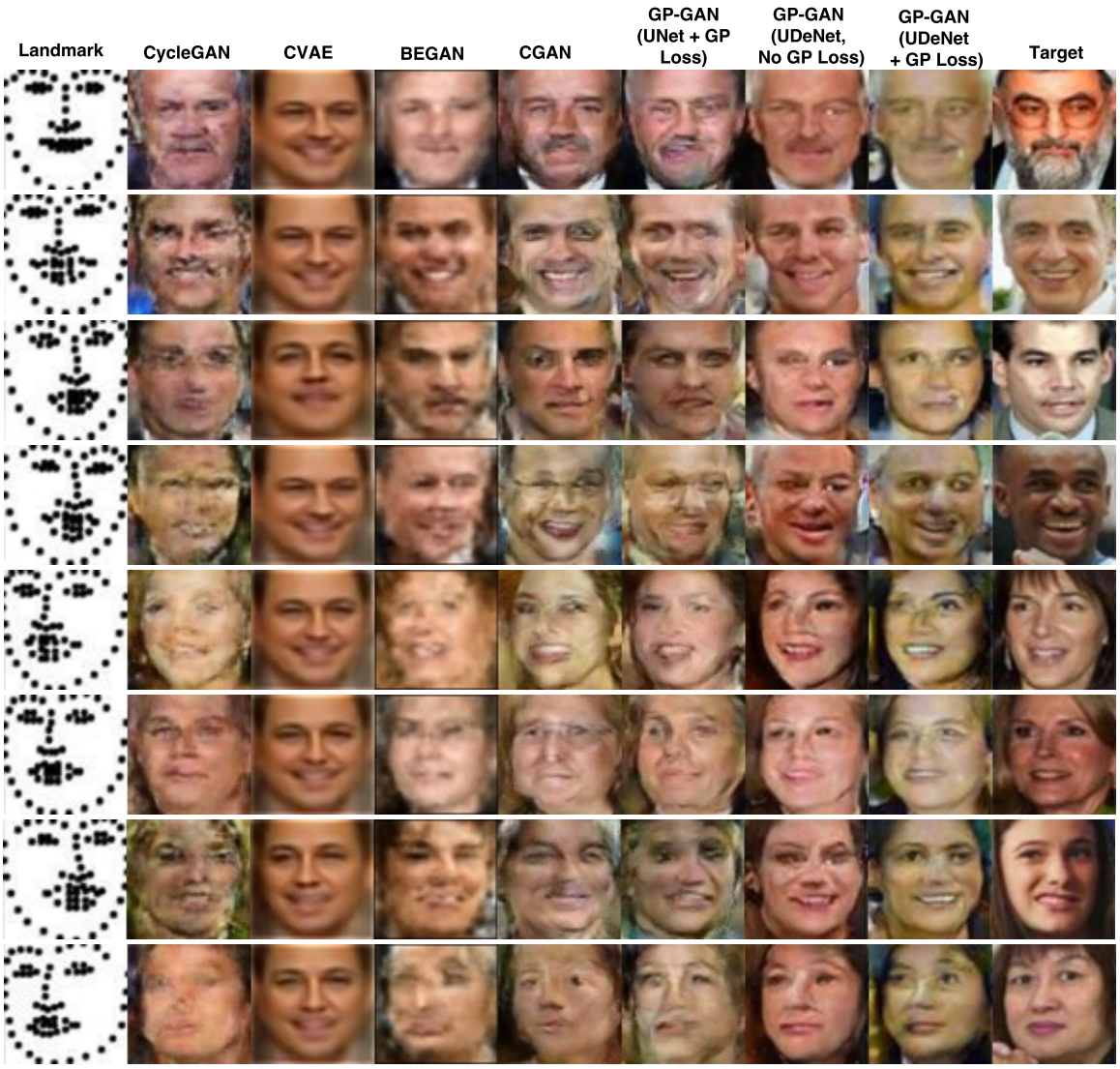


Figure 8.2: Sample qualitative results of synthesis experiments from LFW dataset. The proposed method GP-GAN (UDeNet + GP Loss) achieves more realistic synthesis compared to the other methods (CycleGAN, CVAE, BEGAN, CGAN) and the baseline methods from the ablation study: GP-GAN (UNet+GP Loss), GP-GAN (UDeNet+ No GP Loss).

the corresponding real image and is defined as

$$L_1 = ||G(\hat{x}) - x||_1 \tag{8.5}$$

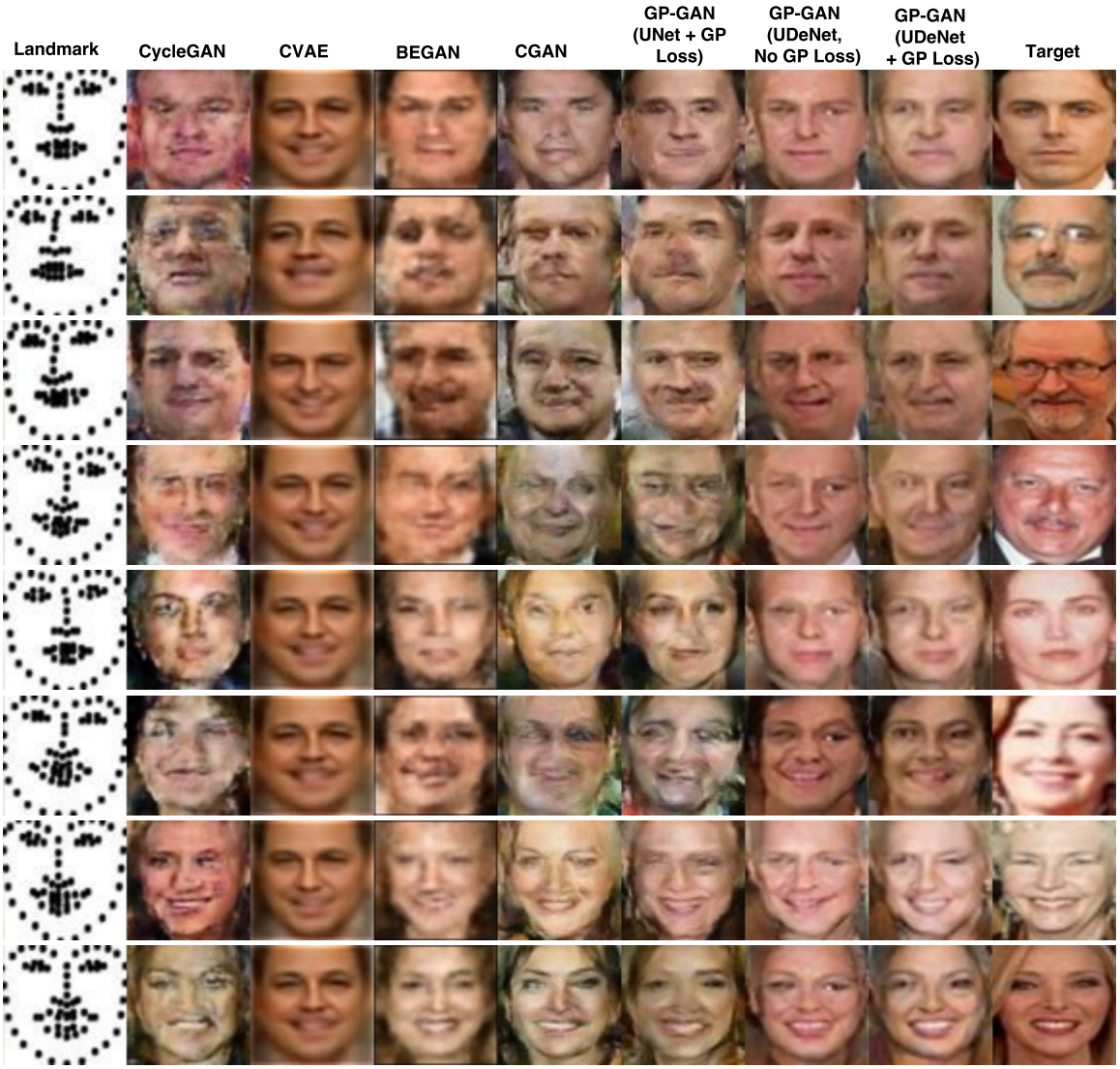


Figure 8.3: Sample qualitative results of synthesis experiments from CASIA WebFace dataset. The proposed method GP-GAN (UDeNet + GP Loss) achieves more realistic synthesis compared to the other methods (CycleGAN, CVAE, BEGAN, CGAN) and the baseline methods from the ablation study: GP-GAN (UNet+GP Loss), GP-GAN (UDeNet+ No GP Loss).

Table 8.1: Quantitative comparison of gender recognition accuracy (%) for various methods.

	LM (D)	LM (A)	CycleGAN	CVAE	BEGAN	CGAN	GP-GAN (UNet+GP-Loss)	GP-GAN (UDeNet,No GP Loss)	GP-GAN (UDeNet+GP-Loss)
LFW	78.0 ± 1.9	79.8 ± 2.4	81.8 ± 1.1	80.3 ± 2.0	84.4 ± 1.9	86.3 ± 2.5	91.1 ± 1.1	91.7 ± 1.6	93.1 ± 1.2
CASIA	61.0 ± 11.8	61.7 ± 13.6	64.8 ± 3.3	62.0 ± 4.1	67.8 ± 5.0	70.4 ± 5.5	73.2 ±3.9	76.7 ± 4.3	78.4 ± 4.1

8.2 Experiments and Evaluations

In this part, experimental settings and evaluation of the proposed method are discussed in detail. We present the qualitative and quantitative results of the synthesis experiment. The quantitative performance is measured using gender recognition rates. Results are compared with four state-of-the-art generative models: Conditional GAN [54], Cycle GAN [119], CVAE [41] [18] and adopted BEGAN[2] in addition to two baseline methods (a) GP-GAN using U-Net generator with GP-Loss, and (b) GP-GAN using UDeNet generator without GP-Loss. The baseline comparisons are performed to demonstrate the improvements achieved by the gender preserving loss and UDeNet components. Also, we demonstrate that the use of synthesis using GP-GAN accentuates gender information present in landmarks by comparing gender recognition rates with methods that directly compute these rates from landmark points [179]. Furthermore, we conduct an experiment to evaluate the data augmentation capabilities of the synthesis method.

8.2.1 Preprocessing and training details

Prior to preforming these experiments, all images in both datasets are fed through a preprocessing pipeline. First, MTCNN [109] is employed for detecting face bounding boxes

[2]https://github.com/taey16/pix2pixBEGAN.pytorch

which are further used to crop the faces followed by landmark key point detection using TCDCN algorithm [185]. Pairs of these detected landmarks and faces are used for training the proposed method. Since we consider this problem as an image-to-image translation, the input landmark is encoded using a heatmap (similar to [198]) as shown in Fig. 8.1 which is a created by imposing a 2D Gaussian with standard deviation of 0.2 at every landmark location on a blank image like could counting work [199]. Note that the cropped face images are resized to 64×64.

The proposed network is trained on a single TitanX GPU for approximately 10 hours (200 epochs). A learning rate of 2×10^{-4} is used for G and D. For perceptual network, the input images are resized to a size of 224×224. The learning rate is decayed by a factor of 2×10^{-6} for every epoch after 100 epochs. The weights λ_A, λ_P and λ_C are set equal to 100, 1 and 1, respectively.

For learning the parameters of the proposed method and baselines, training set from the LFW official deep funneling aligned dataset [200] [108] is used. It contains 5749 identities, and 13233 images. The official training, validating and testing View 1 was used for this experiment. After detection and crop procedure, we are left with 3757 images in the training set and 1615 images in the test set. The trained network is evaluated on the LFW test set and a subset of CASIA-Webface dataset [201]. The test subset for CASIA-Webface is constructed by randomly selecting 1000 male and 1000 female face images. Note that, in order to demonstrate the generalization performance, the proposed network is trained using only the LFW training set and evaluated on the LFW test set and the CASIA-

Webface dataset.

8.2.2 Results

Fig. 8.2 and Fig. 8.3 show sample results of reconstruction using various methods on the LFW and CASIA datasets, respectively. The landmark image is used as the input for all the methods except CVAE [41] [18]. For CVAE, the inputs are original image and normalized landmark locations as the attributes. It can be clearly observed that Conditional GAN [54], Cycle GAN [119] and BEGAN [166] are unable to reconstruct visually coherent faces. Though CVAE is able to generate visually appropriate faces, they fail to preserve the gender information. Since their network implements an auto-encoder like architecture and uses pixel-wise Euclidean measure, the output is often blurry, due to which gender classification becomes very difficult. GP-GAN using UDeNet generator without GP-Loss is able to generate perceptually better results as compared to GP-GAN using UNet generator with GP-Loss demonstrating the superior performance obtained using the novel combination of UNet and DenseNet architectures. The proposed method GP-GAN (UDeNet and GP-Loss) outperforms all existing and baseline methods. It may be argued that identity information is lost during the reconstruction process, however, note that the goal of the proposed method is not to capture the exact mapping between landmarks and corresponding faces. Instead, the idea is to explore generation of visually coherent faces from landmark keypoints which can further assist in data augmentation and other tasks.

As discussed earlier, the quantitative performance is measured in terms of gender recog-

nition rates and it is shown in Table 8.1. Gender recognition rates for the synthesized are calculated using the LBP features [202] and a linear SVM classifier that is trained using the LFW training set, whereas the recognition rates for landmarks, LM(D) and LM(A), are calculated using the distance and angle methods described in [179]. Note that the gender recognition is performed based only on landmark keypoints considering that the corresponding face images are unavailable and hence recent state-of-the-art gender recognition methods cannot be used for comparison as they operate on actual face images rather than only on facial landmarks. Similar to the observations made using visual comparisons, it can be found from the quantitative results that, gender recognition rates improve in general using the generative models as compared to the landmark-based methods.

With respect to the baseline comparisons, it can be observed that GP-GAN using UDeNet generator without GP-Loss outperforms GP-GAN using UNet generator with GP-Loss in spite of the fact that GP-Loss is not used, thus indicating the effectiveness of UDeNet architecture. Furthermore, the proposed method GP-GAN (UDeNet with GP-Loss) outperforms all existing baseline methods by a large margin in terms of gender recognition rates. This indicates that the proposed synthesis method can be used to generate face images from just facial landmarks while retaining gender information present in these landmarks.

In addition, we conducted a face synthesis experiment to verify if the proposed method can be used for data augmentation. In this experiment, we manipulate the landmark of a face (for instance, modify mouth open to mouth close) and use this landmark to synthesize a face using generator G. Sample results for this experiment are shown in Fig 8.4. It can be

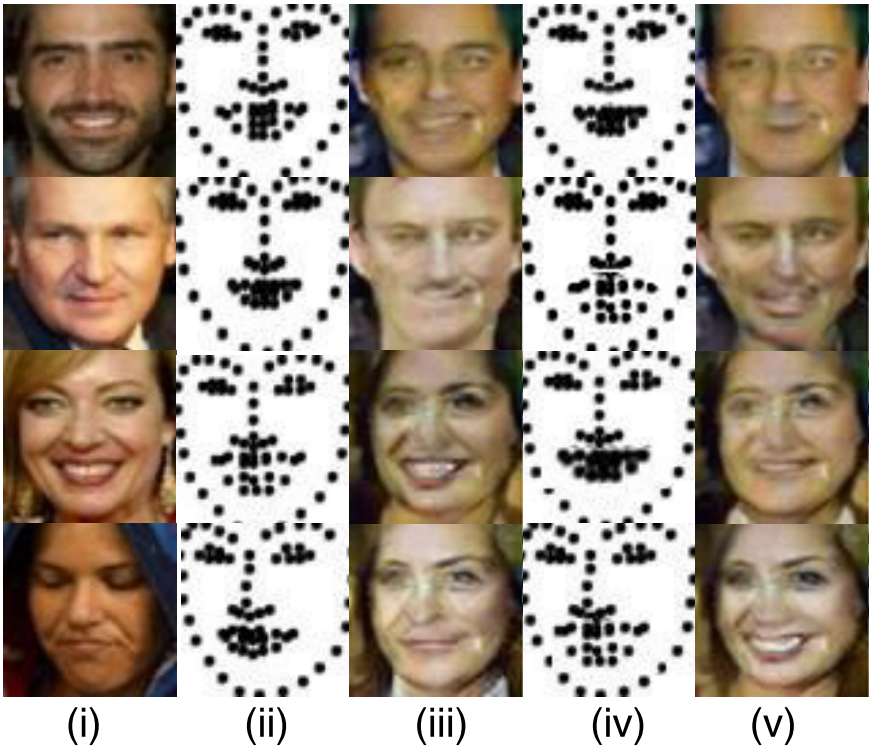


Figure 8.4: Results of experiment for dataset augmentation where landmark corresponding to a face is modified and used for synthesis. We are able to generate new samples while preserving gender information. (i) Original face image. (ii) Landmark corresponding to original face. (iii) Synthesized face from original landmark. (iv) Landmark obtained after manipulating original landmark. (v) Synthesized face image using manipulated landmark.

seen that, the generator G is able to synthesize realistic faces from the modified landmarks while reflecting this modification in the synthesized face. Additionally, the gender attribute is also retained. Based on these experiments, we can conclude that the proposed method is able successfully generate face samples which can be used for data augmentation for other facial analysis tasks.

8.3 Summary

We explored the problem of synthesizing faces from landmarks points using the recently introduced generative models. The aim of this project was to demonstrate that information (especially gender) present in the landmark keypoints can be accentuated using synthesis models while generating realistic images. The proposed network is based on the generative adversarial networks and is guided by perceptual loss and a novel gender preserving loss. Further, we propose a novel generator based on UNet and DenseNet architectures. Evaluations are performed on two popular datasets, LFW and CASIA-Webface, and the results are compared with recent state-of-the-art generative methods. It is clearly demonstrated that the proposed method achieves significant improvements in terms of visual quality and gender recognition. Additionally, we conducted a face synthesis experiment to demonstrate that the proposed generative method can be used as a data augmentation technique.

Chapter 9

Discussion and Future Work

In this thesis, we developed several facial image synthesis problem using deep generative networks.

Specifically, we proposed the Att2Sk2Face model for face image synthesis from visual attribute via sketch. The proposed network was shown to synthesize photo-realistic facial images conditioned on the visual attributes input. We proposed a two-stage solution to this problem: (1) attribute-to-sketch (2) sketch-to-image. For each part, the Sketch Generator Network and Face Generator Network are trained by leveraging the advances of hierarchical network architecture. Additionally, we develop a multimodal face synthesis model (Att2MFace) which simultaneously generates face images in different modalities via one given visual attributes. The developed Att2MFace consists of a single generator and discriminator. The multimodal stretch-in/stretch-out modules are introduced in generator/discriminator to learn the discriminability between real/synthesized multimodal

images. An auxiliary attribute estimator is accompanying with the discriminator to learn the matching between the synthetic images and visual attributes. The progressive training strategy is adopted here to enhance the image quality between the synthesized multimodal images.

Also, several thermal-to-visible face synthesis models are developed in this thesis. For instance, we introduced the thermal-to-visible face synthesis network guided by the visual attributes for heterogeneous face verification. Rather than image-level information, the proposed Multi-AP-GAN also leverages the abstract semantic information provided by the visual attributes. The attributes are extracted by a pre-trained attribute estimator and the extracted attributes are fused with the image latent representation with multimodal compact bilinear pooling. Additionally, the training is regularized by the multi-scale resolution target images for better synthesis. Various experiments on different datasets demonstrated the effectiveness of the proposed scheme.

Another self-attention guided generative adversarial network is proposed for thermal to visible face synthesis. With the help of self-attention module better synthesis quality is achieved. Additionally, the averaged features from the synthesized image and the input images are used for obtaining better performance on heterogeneous verification.

In addition, a heterogeneous face frontalization model is proposed in this thesis. Given the arbitrary yaw pose face images in thermal domain, the proposed model is able to generate identity-preserving frontal face image in the visible domain. To achieve this, the proposed frontalization model contains two streams: cross-spectrum frontalization and do-

main agnostic learning. The cross-spectrum frontalization stream aims to generate the frontal visible face from learned feature representation while the domain agnostic learning stream aims to learn the feature to be robust to the domain discrepancy. Also, we show that learning domain agnostic feature is not enough for identity-preserving synthesis and learning discriminative feature is also important .To achieve this, a contrasitive loss is leveraged by a proposed two-path training strategy.

Finally, we demonstrate the capability of generative adversarial networks in another kind of abstract semantic representation: landmarks. In this work, we demonstrated the information (especially gender) in landmarks can be recovered by reconstructing the face images. A novel gender preserving loss is presented with the inspiration of perceptual loss. Additionally, a dense block based UNet architecture is developed and extensive experiments clearly demonstrate the significant improvements by the proposed method in both visual quality and gender recognition.

In the future, we hope to address the following challenges in face synthesis. First, we would like to learn a text-to-face synthesis model instead of visual attributes-to-face. This can be even more challenging because learning discriminative text representation is very difficult. Secondly, instead of a single image frame, we would like to tackle the problem of visual attribute guided video clip (a series of consecutive image frames) synthesis. The learned video clip could be more applicable in law enforcement and entertainment. Additionally, we would like to tackle the problem of multi-style multimodal face synthesis from visual attributes. Generating multimodal face images in different styles: illumination, pose,

etc. could improve the diversity of the synthesized images.

Bibliography

[1] S. Hu, J. Choi, A. L. Chan, and W. R. Schwartz, "Thermal-to-visible face recognition using partial least squares," *JOSA A*, vol. 32, no. 3, pp. 431–442, 2015.

[2] B. Klare and A. K. Jain, "Heterogeneous face recognition: Matching nir to visible light images," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 1513–1516.

[3] J. Choi, S. Hu, S. S. Young, L. S. Davis, "Thermal to visible face recognition," in *Proc.SPIE*, 2012, pp. 8371 – 8371 – 10.

[4] T. Bourlai, A. Ross, C. Chen, L. Hornak, "A study on using mid-wave infrared images for face recognition," vol. 8371, 2012, pp. 8371 – 8371 – 13.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[6] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and ap-

proximate inference in deep generative models," in *International Conference on Machine Learning*, 2014.

[7] H. Larochelle and I. Murray, "The neural autoregressive distribution estimator," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 29–37.

[8] A. V. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1747–1756. [Online]. Available: http://proceedings.mlr.press/v48/oord16.html

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] T. Xiao, J. Hong, and J. Ma, "Elegant: Exchanging latent encodings with gan for transferring multiple face attributes," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 168–184.

[11] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9243–9252.

BIBLIOGRAPHY

[12] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[13] Y. Yin, S. Jiang, J. Robinson, and Y. Fu, "Dual-attention gan for large-pose face frontalization," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pp. 24–31.

[14] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8398–8406.

[15] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2439–2448.

[16] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2492–2501.

[17] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 217–233.

[18] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 776–791.

[19] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, pp. 1060–1069. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045390.3045503

[20] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *International Conference on Computer Vision*, 2017, pp. 5907–5915.

[21] ——, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, 2019.

[22] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6199–6208.

[23] D. Bang and H. Shim, "Resembled generative adversarial networks: Two domains with similar attributes," in *29th British Machine Vision Conference, BMVC 2018*, 2019.

[24] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in neural information processing systems*, 2016, pp. 469–477.

[25] X. Di and V. M. Patel, "Multimodal face synthesis from visual attributes," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 427–439, 2021.

[26] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu, "Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces," in *IEEE International Joint Conference on Biometrics (IJCB)*, Oct 2017, pp. 100–107.

[27] B. S. Riggan, N. J. Short, S. Hu, and H. Kwon, "Estimation of visible spectrum faces from polarimetric thermal faces," in *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*. IEEE, 2016, pp. 1–7.

[28] B. S. Riggan, N. J. Short, and S. Hu, "Thermal to visible synthesis of face images using multiple regions," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[30] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *International Conference on Computer Vision*, December 2015.

[31] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[32] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016.

[33] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 317–326.

[34] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks," *International Journal of Computer Vision: Special Issue on Deep Learning for Face Analysis*, 2019.

[35] X. Di, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Multi-scale thermal to visible face verification via attribute guided synthesis," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 266–280, 2021.

[36] K. Mallat and J.-L. Dugelay, "A benchmark database of visible and thermal paired face images across multiple variations," in *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2018, pp. 1–5.

[37] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, and X. Yuan, "A comprehensive database for benchmarking imaging systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 509–520, 2020.

[38] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Hk99zCeAb

[39] S. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 348–353.

[40] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[41] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015, pp. 3483–3491.

[42] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *International Conference on Machine Learning*, 2016.

[43] E. L. Denton, S. Chintala, R. Fergus *et al.*, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Advances in neural information processing systems*, 2015, pp. 1486–1494.

[44] A. Dosovitskiy, J. T. Springenberg, M. Tatarchenko, and T. Brox, "Learning to generate chairs, tables and cars with convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 692–705, 2017.

[45] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[46] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," *International Conference for Learning Representations*, 2017.

[47] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *International Conference for Learning Representations*, 2017.

[48] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," *International Conference for Learning Representations*, 2017.

[49] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," 2015.

[50] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary

classifier gans," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2642–2651.

[51] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Cvae-gan: Fine-grained image generation through asymmetric training," in *International Conference on Computer Vision*, 2017.

[52] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 627–635.

[53] Z. Lu, T. Hu, L. Song, Z. Zhang, and R. He, "Conditional expression synthesis with face parsing transformation," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 1083–1091.

[54] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition*, 2017.

[55] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *Computer Vision and Pattern Recognition*, 2014.

[56] P. Li, Y. Hu, R. He, and Z. Sun, "Global and local consistent wavelet-domain age synthesis," *IEEE Transactions on Information Forensics and Security*, 2019.

[57] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional

neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3677–3685.

[58] S. Banerjee, W. Scheirer, K. Bowyer, and P. Flynn, "On hallucinating context and background pixels from a face mask using multi-scale gans," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 300–309.

[59] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1316–1324.

[60] T. Zhang, A. Wiliem, S. Yang, and B. Lovell, "Tv-gan: Generative adversarial network based thermal to visible face recognition," in *2018 International Conference on Biometrics (ICB)*, 2018, pp. 174–181.

[61] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Adversarial cross-spectral face completion for nir-vis face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[62] A. Yu, H. Wu, H. Huang, Z. Lei, and R. He, "Lamp-hq: A large-scale multi-pose high-quality database and benchmark for nir-vis face recognition," *International Journal of Computer Vision*, pp. 1–17, 2021.

[63] X. Di, H. Zhang, and V. M. Patel, "Polarimetric thermal to visible face verification via attribute preserved synthesis," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–10.

[64] X. Di, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Polarimetric thermal to visible face verification via self-attention guided synthesis," in *2019 International Conference on Biometrics (ICB 2019)*, 2019.

[65] T. de Freitas Pereira, A. Anjos, and S. Marcel, "Heterogeneous face recognition using domain specific units," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1803–1816, 2018.

[66] A. Litvin, K. Nasrollahi, S. Escalera, C. Ozcinar, T. B. Moeslund, and G. Anbarjafari, "A novel deep network architecture for reconstructing rgb facial images from thermal for face recognition," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 25 259–25 271, 2019.

[67] O. Nikisins, K. Nasrollahi, M. Greitans, and T. B. Moeslund, "Rgb-d-t based face recognition," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 1716–1721.

[68] K. Mallat, N. Damer, F. Boutros, A. Kuijper, and J.-L. Dugelay, "Cross-spectrum thermal to visible face recognition based on cascaded image synthesis," in *ICB 2019, 12th IAPR International Conference On Biometrics, 4-7 June, Crete, Greece*, Crete, GRÈCE, 06 2019.

[69] N. Damer, F. Boutros, K. Mallat, F. Kirchbuchner, J.-L. Dugelay, and A. Kuijper, "Cascaded generation of high-quality color visible face images from thermal captures," *arXiv preprint arXiv:1910.09524*, 2019.

[70] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 768–783.

[71] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional cyclegan," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 282–297.

[72] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4295–4304.

[73] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 787–796.

[74] L. A. Jeni and J. F. Cohn, "Person-independent 3d gaze estimation using face frontalization," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 87–95.

[75] J. Zhao, L. Xiong, Y. Cheng, Y. Cheng, J. Li, L. Zhou, Y. Xu, J. Karlekar, S. Pranata,

S. Shen, J. Xing, S. Yan, and J. Feng, "3d-aided deep pose-invariant face recognition," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, ser. IJCAI'18. AAAI Press, 2018, p. 11841190.

[76] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3871–3879.

[77] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3d morphable model learnt from 10,000 faces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5543–5552.

[78] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3d morphable models," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 233–254, 2018.

[79] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Learning a high fidelity pose invariant model for high-resolution face frontalization," in *NeurIPS*, 2018, pp. 2872–2882. [Online]. Available: http://papers.nips.cc/paper/7551-learning-a-high-fidelity-pose-invariant-model-for-high-resolution-face-frontalization

[80] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1415–1424.

[81] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3990–3999.

[82] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing *et al.*, "Towards pose invariant face recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2207–2216.

[83] Y. Wei, M. Liu, H. Wang, R. Zhu, G. Hu, and W. Zuo, "Learning flow-based feature warping for face frontalization with illumination inconsistent supervision," in *Proceedings of the European Conference on Computer Vision*, 2020.

[84] Z. Zhang, X. Chen, B. Wang, G. Hu, W. Zuo, and E. R. Hancock, "Face frontalization using an appearance-flow-based convolutional neural network," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2187–2199, 2018.

[85] S. Jiang, Z. Tao, and Y. Fu, "Geometrically editable face image translation with adversarial networks," *IEEE Transactions on Image Processing*, 2021.

[86] P. Li, X. Wu, Y. Hu, R. He, and Z. Sun, "M2fpa: a multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 043–10 051.

[87] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, "Describable visual attributes

for face verification and image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1962–1977, Oct 2011.

[88] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? a survey on soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 441–467, March 2016.

[89] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *International Conference on Computer Vision*, Dec 2015, pp. 3730–3738.

[90] N. Kumar, P. Belhumeur, and S. Nayar, "Facetracer: A search engine for large collections of images with faces," in *European conference on computer vision*. Springer, 2008, pp. 340–353.

[91] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1637–1644.

[92] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.

[93] E. M. Rudd, M. Günther, and T. E. Boult, "Moon: A mixed objective optimization

network for the recognition of facial attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 19–35.

[94] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris, "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5334–5343.

[95] M. Günther, A. Rozsa, and T. E. Boult, "Affact: Alignment-free facial attribute classification technique," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 90–99.

[96] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *European Conference on Computer Vision*, 2016.

[97] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," *arXiv preprint arXiv:1706.08033*, 2017.

[98] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[99] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[100] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 83–90.

[101] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Hk99zCeAb

[102] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[103] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[104] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," in *Advances in Neural Information Processing Systems*, 2017, pp. 6594–6604.

[105] X. Di, H. Zhang, and V. M. Patel, "Polarimetric thermal to visible face verification via attribute preserved synthesis," in *Biometrics: Theory, Applications, and Systems (BTAS), IEEE International Conference on*. IEEE, 2018.

[106] X. Di, V. A. Sindagi, and V. M. Patel, "Gp-gan: Gender preserving gan for syn-

thesizing faces from landmarks," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1079–1084.

[107] X. Di and V. M. Patel, "Face synthesis from visual attributes via sketch using conditional vaes and gans," *arXiv preprint arXiv:1801.00077*, 2017.

[108] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller, "Learning to align from scratch," in *Advances in Neural Information Processing Systems*, 2012.

[109] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[110] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[111] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[112] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[113] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are gans created equal? a large-scale study," in *Advances in neural information processing systems*, 2018, pp. 700–709.

[114] Y. Wang, A. Dantcheva, and F. Bremond, "From attributes to faces: a conditional generative network for face generation," in *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sep. 2018, pp. 1–5.

[115] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5907–5915.

[116] X. Di and V. M. Patel, "Facial synthesis from visual attributes via sketch using multiscale generators," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 1, pp. 55–67, 2019.

[117] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–189.

[118] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett,

Eds. Curran Associates, Inc., 2017, pp. 700–708. [Online]. Available: http://papers.nips.cc/paper/6672-unsupervised-image-to-image-translation-networks.pdf

[119] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *International Conference on Computer Vision*, 2017.

[120] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *International Conference on Computer Vision*, 2017.

[121] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[122] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 83–90.

[123] X. Wu, H. Huang, V. M. Patel, R. He, and Z. Sun, "Disentangled variational representation for heterogeneous face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9005–9012.

[124] H. Kong, J. Zhao, X. Tu, J. Xing, S. Shen, and J. Feng, "Cross-resolution face recognition via prior-aided face hallucination and residual knowledge distillation," *arXiv preprint arXiv:1905.10777*, 2019.

[125] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan, and J. Feng, "Dual-agent gans for photorealistic and identity preserving profile face synthesis," in *Advances in neural information processing systems*, 2017, pp. 66–76.

[126] J. Cao, H. Huang, Y. Li, J. Liu, R. He, and Z. Sun, "Biphasic learning of gans for high-resolution image-to-image translation," *arXiv preprint arXiv:1904.06624*, 2019.

[127] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[128] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "Dual variational generation for low shot heterogeneous face recognition," in *Advances in Neural Information Processing Systems*, 2019, pp. 2674–2683.

[129] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3d-aided dual-agent gans for unconstrained face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 10, pp. 2380–2394, 2018.

[130] X. Di, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Polarimetric thermal to visible face verification via self-attention guided synthesis," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[131] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep

convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[132] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2642–2651.

[133] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[134] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[135] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in neural information processing systems*, 2017, pp. 5767–5777.

[136] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *International Conference on Machine Learning*, 2017.

[137] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.

[138] X. Mao and Q. Li, "Unpaired multi-domain image generation via regularized conditional gans," in *27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2018)*. International Joint Conferences on Artificial Intelligence, 2018, pp. 2553–2559.

[139] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8188–8197.

[140] S. Hu, N. J. Short, B. S. Riggan, C. Gordon, K. P. Gurton, M. Thielke, P. Gurram, and A. L. Chan, "A polarimetric thermal database for face recognition research," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 119–126.

[141] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[142] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Advances in Neural Information Processing Systems*, 2017.

[143] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 35–51.

[144] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J. C. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa, "Deep learning for understanding faces: Machines may be just as good, or better, than humans," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 66–83, Jan 2018.

[145] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[146] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[147] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[148] J. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–9.

[149] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning

framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2017.

[150] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[151] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[152] N. Short, S. Hu, P. Gurram, K. Gurton, and A. Chan, "Improving cross-modal face recognition using polarimetric imaging," *Optics letters*, vol. 40, no. 6, pp. 882–885, 2015.

[153] F. Nicolo and N. A. Schmid, "Long range cross-spectral face recognition: matching swir against visible light images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1717–1726, 2012.

[154] J. Lezama, Q. Qiu, and G. Sapiro, "Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 6807–6816.

[155] T. Bourlai, N. Kalka, A. Ross, B. Cukic, and L. Hornak, "Cross-spectral face verifi-

cation in the short wave infrared (swir) band," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 1343–1347.

[156] T. Bourlai and L. A. Hornak, "Face recognition outside the visible spectrum," *Image and Vision Computing*, vol. 55, pp. 14 – 17, 2016.

[157] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 239–247.

[158] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=B1QRgziT-

[159] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[160] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[161] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.

[162] K. Mallat and J.-L. Dugelay, "A benchmark database of visible and thermal paired

face images across multiple variations," in *International Conference of the Biometrics Special Interest Group, BIOSIG 2018, Darmstadt, Germany, September*, ser. LNI. GI / IEEE, pp. 199 – 206.

[163] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[164] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 5188–5196.

[165] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networkss," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[166] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

[167] B. S. Riggan, N. J. Short, and S. Hu, "Optimal feature learning and discriminative framework for polarimetric thermal to visible face recognition," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–7.

[168] B. S. Riggan, N. J. Short, M. S. Sarfraz, S. Hu, H. Zhang, V. M. Patel, S. Rasnayaka, J. Li, T. Sim, S. M. Iranmanesh, and N. M. Nasrabadi, "Icme grand challenge re-

sults on heterogeneous face recognition: Polarimetric thermal-to-visible matching," in *2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2018, pp. 1–4.

[169] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 53–69, May 2015.

[170] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.

[171] T. de Freitas Pereira, A. Anjos, and S. Marcel, "Heterogeneous face recognition using domain specific units," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1803–1816, 2019.

[172] S. Liu, J. Yang, C. Huang, and M.-H. Yang, "Multi-objective convolutional learning for face labeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3451–3459.

[173] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates,

Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf

[174] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[175] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.

[176] D. Poster, M. Thielke, R. Nguyen, S. Rajaraman, X. Di, C. N. Fondje, V. M. Patel, N. J. Short, B. S. Riggan, N. M. Nasrabadi, and S. Hu, "A large-scale, time-synchronized visible and thermal face dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021, pp. 1559–1568.

[177] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani *et al.*, "A comprehensive database for benchmarking imaging systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 509–520, 2018.

[178] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

BIBLIOGRAPHY

[179] D. Cao, C. Chen, M. Piccirilli, D. Adjeroh, T. Bourlai, and A. Ross, "Can facial metrology predict gender?" in *2011 International Joint Conference on Biometrics (IJCB)*. IEEE, 2011, pp. 1–8.

[180] T. Wu, P. Turaga, and R. Chellappa, "Age estimation and face verification across aging using landmarks," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1780–1788, 2012.

[181] Q. Sun, L. Ma, S. J. Oh, L. V. Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," in *CVPR*, 2018.

[182] W. Wang, X. Alameda-Pineda, D. Xu, E. Ricci, and N. Sebe, "Every smile is unique: Landmark-guided diverse smile generation," in *CVPR*, 2018.

[183] J. C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–9.

[184] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J. C. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa, "Deep learning for understanding faces: Machines may be just as good, or better, than humans," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 66–83, Jan 2018.

[185] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *ECCV*. Springer, 2014, pp. 94–108.

[186] S. Taheri, P. Turaga, and R. Chellappa, "Towards view-invariant expression analysis using analytic shape manifolds," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 306–313.

[187] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[188] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," *arXiv preprint arXiv:1609.03126*, 2016.

[189] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *arXiv preprint arXiv:1701.05957*, 2017.

[190] L. Wang, V. A. Sindagi, and V. M. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," *CoRR*, vol. abs/1710.10182, 2017. [Online]. Available: http://arxiv.org/abs/1710.10182

[191] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3194–3203.

[192] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2729–2736.

[193] A. Kumar, R. Ranjan, V. Patel, and R. Chellappa, "Face alignment by local deep descriptor regression," *arXiv preprint arXiv:1601.07950*, 2016.

[194] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3476–3483.

[195] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," *arXiv preprint arXiv:1612.05424*, 2016.

[196] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Computer Vision and Pattern Recognition*, 2017.

[197] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections," in *Advances in Neural Information Processing Systems.*, 2016.

[198] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *NIPS*, 2017.

[199] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using

contextual pyramid cnns," in *IEEE International Conference on Computer Vision*, 2017.

[200] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep.

[201] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[202] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.